当代经济管理权威名著

·珍藏版·

世界经济管理文库

经济学：原理与政策

威廉 J. 鲍默尔
艾伦 S. 布兰德 /著



Economics: Principles and Policy

William J. Baumol　Alan S. Blinder

SEVENTH EDITION
英文版·第⑦版

机械工业出版社　DRYDEN

当代经济管理权威读著

经济学：原理与政策

（英文版·第7版）

Economics: Principles and Policy

(SEVENTH EDITION)

威廉 J. 鲍默尔
(William J. Baumol) /著
艾伦 S. 布兰德
(Alan S. Blinder)

The Dryden Press
Harcourt Brace College Publishers

Harcourt Brace & Company
Asia Pte Ltd

机械工业出版社

Reprint ISBN: 0030211328

本书版权登记号：图字：01-98-0940

图书在版编目（CIP）数据

经济学：原理与政策：第7版：英文/（美）鲍默尔（Baumol, W. J.），（美）布兰德（Blinder, A. S.）著.-影印版.-北京：机械工业出版社，1998.7
（当代经济管理权威名著）
ISBN 7-111-06474-7

Ⅰ.经… Ⅱ.①鲍… ②布… Ⅲ.经济学-英文 Ⅳ.F01

中国版本图书馆 CIP 数据核字（98）第12579号

出 版 人：马九荣（北京百万庄大街22号 邮政编码100037）
责任编辑：刘露明
审 读 人：陈 跃
昌平环球印刷厂印刷·新华书店北京发行所发行
1998年6月第1版第1次印刷
787mm×1092mm 1/16·61.75印张
印 数：0001-2000册
定 价：99.00元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

致读者(代序)

这是一本供初学者使用的经济学入门书,也是一本供经济学的老学生温故而知新的经济学手册。它可以作为教材使用,也可以供自学之用。

当你在决定是否购买这本书之前,你需要考虑的是:你能从这本书中得到什么收益?这种收益是否能补偿你为此付出的成本?(这种成本不仅包括你支付的货币,而且包括你读这本书所要花费的时间。一般来说,读书的时间成本远远大于货币成本,但许多人常常忘记了这一点,所以他们在决定是否买一本书时,过多地考虑了书价的高低。)下面的问题或许有助于你作出决策。

一、为什么要学习经济学

经济学是经济学家提供给社会大众的一种改进生活、认识世界的武器。或许你并不想做一名经济学家,但即使如此,你仍然应该学点经济学。

首先,学习经济学有助于你作出更好的个人决策。在你的一生中,你需要作出各种各样的经济决策。比如说,在即将完成高中学业的时候,你需要决定是否去上大学?在大学毕业的时候,你需要决定是继续在国内读研究生,还是出国留学,或者去工作?在工作之后,你要决定如何花费你的收入:多少用于现在的消费?多少用于储蓄?如何投资你的储蓄?是买股票还是存在银行?或许有一天你成了一个企业的老板或经理,此时,你需要决定你的企业应该生产什么产品?卖什么样的价格?在什么媒体上作广告?招收什么样的人员?提拔谁当你的助手?如此等等,不一而举。为什么决策是重要的?因为你的资源是有限的——你的时间有限,收入也有限。如果你参加工作,就可能没有时间上大学;如果你把钱用于买房子,就可能没有钱再来买汽车。所以你必须在各种竞争性的需求之间分配你有限的资源。进一步,更为麻烦的是,你的决策常常是在不确定的情况下作出的。比如说,当你选择学无线电专业的时候,你并不确定当你毕业的时候,这个专业的就业前景如何。为了避免决策的失误,你需要一些理论的指导。经济学是有关个人选择的科学,学习经济学有助于你作出更好的决策。明白了这一点,你就明白了为什么经济学是西方大学里听众最多的选修课。

其次,学习经济学有助于你理解你生活于其间的世界是如何运转的。你的生活状况不仅取决于你自己的决策,而且依赖于其他人的决策,以及周围环境的变化。理解你周围的世界如何运行,自然有助于改进你的决策。你可能为生活中的许多事情感到惊奇。比如说,当你想买一台电视机的时候,只要支付必要的价格,就可以把它从商场搬回家;当你在大街上走得饥肠辘辘的时候,走进一家饭馆就可以指挥别人给你上菜,而无须回家自己做饭。而事实上,你事前并没有告诉电视机的生产厂家为你生产一台电视机,也没有通知饭馆的老板为你准备饭菜。那么,是什么因素使你得到你想要的东西?经济学是有关人们之间的决策如何相互作用的科学。学了经济学,你就可以明白市场这只"看不见的手"如何使自利的个人为大家服务;你也可以明白,为什么中国电信的服务那么差,而收费却那么高;为什么北京的"面的"司机有拒载现象,而奥迪车司机却不拒载;为什么一个流行歌手演出一晚上可以赚好几万元,而一个建筑个人一个

月却只能赚几百元；为什么利率一上升股票价格就下跌；为什么中国经济增长这么快，而下岗工人却越来越多；为什么国有企业设备那么好，而就是竞争不过民营企业；为什么那么多国有企业的厂长经理一生勤勤恳恳，一到 59 岁就犯贪污罪……。

再次，学习经济学有助于你理解政府政策的优与劣。每个社会都离不开政府，学习了经济学，你会明白我们为什么需要政府、什么是政府应该干的，什么是政府不应该干的。我们需要政府，是因为在存在诸如外部性、公共产品这样的场合，依靠市场不能达到资源的有效配置。比如说，如果没有政府的干预，自私自利的企业家也许会使你喝太多的污水；如果没有政府，私人部门也许不会提供你诸如路灯这样的公共产品。特别地，我们需要政府来提供市场交易所需要的规则和秩序这样一类公共产品，需要政府保护我们的个人财产和人身安全。但政府对市场的过多干预常常导致供给不足、价格扭曲、资源浪费、垄断横行。政府的政策选择不仅影响整个社会的资源配置效率，而且影响包括你在内的每个公民的福利。所以当你希望政府制定某种政策的时候，你必须谨慎考虑这种政策的不利后果。经济学常识有助于你思考这样的问题。比如说，如果你是一个打工仔，你也许会认为政府应该制定一个"最低工资法"保护你的利益，或者向富人多征税来补贴你，而经济学原理会告诉你，这样做的后果也许是你根本就找不到工作。再比如，作为消费者，你也许会认为政府应该对商品的价格作出限制，而经济学会告诉你，这样做的后果是你也许再也买不到这种商品。常识可能会误导你，经济学会使你不过于迷信政府。

最后，学习经济学可以改进你的思考方式。经济学是一门科学。科学是什么？科学是一种思考问题的方式。如同天文学家通过观测天体现象来归纳天体运行规律一样，经济学家通过观测现实经济现象归纳经济规律。经济学家有自己的语言和思维方式。诸如需求、供给、弹性、消费者剩余、机会成本、比较优势、外部性、信息不对称、均衡等等，是经济学的基本语言。掌握了这些经济学语言，你就可以更好地思考你周围的世界是如何运行的。外行人常常批评经济学家看问题过于简单，而他们忘了，科学的力量就在于把复杂的现象简单化。经济学就象一张指路图，它舍弃了现实中的许多细节，却使你更清楚自己要去的地方。你没有必要成为一位经济学家，但知道经济学是如何思考问题是有益的，至少，你不大容易被蹩脚的经济学家和夸夸其谈的政治家所蒙蔽。无论你今后干什么，你不会后悔自己学过经济学。

二、为什么选择这本书

如果你已经决定学点经济学，接下来要作出的一个决策是选择哪本书入门？你当然希望选择一本最好的教科书。问题是选择好的教科书并不是一件容易的事，因为图书市场充斥着众多的教科书，质量参差不齐。每位作者都会声称自己的书最好，而作为初学者，你不大可能有能力判断哪本好，哪本不好。这就是经济学家讲的信息不对称：买的人不如卖的精。

在这种情况下，最安全的办法当然是选择最权威的经济学家写的最新版的教科书。为什么要选择最权威的经济学家写的教科书？原因不仅是因为最权威的经济学家炉火纯青，对理论的把握最准确，能深入浅出地写出"正经"来；而且因为最权威的经济学家最注重自己的名声，他们既没有必要用出一本新书的办法增加自己的知名度，也不敢为了赚取版税而滥竽充数地出书。这就如同麦当劳不敢卖过时的汉堡包一样。他们的声誉就是质量的代号，他们的名字就是一种品牌。

为什么要选择新版的教科书？因为如同经济生活本身一样，经济学也在不断发展和演化，以用新的理论解释新的经济现象。特别是过去 20 多年里，随着博弈论、信息经济学和理性预期

的引入，经济学可以说发生了一场革命。经济学理论越来越贴近现实，对现实的解释能力越来越强。比如说，20多年前，经济学理论还难以为企业经营决策提供多少理论指导，而今天，博弈论已成为企业战略管理的基础；20多年前，经济学家对公司融资几乎无话可说，而今天，信息经济学已成了公司融资理论的基础；20多年前，经济学家还在迷信于用政府管制来弥补市场的失灵，而今天，政府管制被认为是资源配置无效率的最主要原因之一，从而导致了席卷全球的经济自由化浪潮。或许，我们可以将70年代经济学教科书与90年代经济学教科书的差别，类比于286计算机与586计算机的差别。只有使用最新版的教科书，你才能学到最新的经济学理论。

过去，中国教科书市场上很难看到最权威的经济学家写的最新版的经济学教科书，翻译过来的教科书一般滞后于原版10年以上，而国内经济学家写的"西方经济学"教科书其内容则更为陈旧，更不用说在有色眼镜透视下的扭曲了。可喜的是，近几年来，随着中国的改革开放，经济学图书市场正在发生一场革命性的变化，不仅国外最新版的英文著作在及时引进，而且国内经济学家也出版了一些在内容上与国外同步的经济学教科书。

在此，我愿意向你推荐机械工业出版社最新出版的如下三本教科书：

1. N. Gregory Mankiw 著：The Principles of Economics（经济学原理），1998年第1版。

2. William Baumol 和 Alan Blinder 合著：Economics —— principles and policy（经济学：原理与政策），1997年第7版。

3. Paul Samuelson 和 William Nordhaus 合著：Economics（经济学），1998年第16版。

这三本书都是享誉全球的经济学大师撰写的最新版的经济学入门书，是西方大学里最流行的教科书。

N. Gregory Mankiw 是美国著名的年轻一代经济学家，早年毕业于普林斯顿大学，并在MIT获得博士学位，现为哈佛大学经济学教授，同时兼任著名的智囊机构国家经济研究局（NBER）货币经济学项目主任，波斯顿联邦储备银行和美国国会预算办公室顾问。他的《宏观经济学》是美国最畅销的教科书之一。他是一位多产而富有创新思想的经济学家，在多种一流学术杂志上发表了数十篇开创性的论文，被公认为是新凯恩斯经济学的代表人物之一。他在经济学界的名气比他的年龄大得多。

N. Gregory Mankiw 的《经济学原理》今年一上市，就成为最畅销的教科书，出版商在付印之前就支付了作者100万美元的版税，创下了教科书版税的新纪录。这本教科书的最大特点是它的"学生导向"。如作者自己指出的，他要写的是一本学生喜欢看的经济学教科书。与其他同类教科书相比，本书更多地强调的是经济学原理的应用和政策分析，而不是正式的经济学模型。为此，作者在大部分章节里都提供了案例，以说明经济学原理如何应用于现实经济问题的分析。此外，作者在书中还提供了大量的"新闻摘要"，以使读者懂得什么是生活中的经济学。

William Baumol 和 Alan Blinder 也是两位享誉全球的著名经济学家。Baumol 生于1922年，在美国纽约州立大学获得艺术学学士学位，在英国伦敦经济学院获得经济学博士学位，曾执教于普林斯顿大学，现为纽约州立大学经济学教授，曾担任过美国经济学会主席等多种学术职务。他在产业组织、企业理论、经济增长、反垄断理论等方面作出重要的开创性贡献，发表了500多篇学术论文和20多本著作，其中有些著作被翻译成多种文字。

Blinder 生于1945年，在普林斯顿大学获得经济学学士学位，在MIT获得经济学博士学位，自1979年以来，Blinder 任普林斯顿大学经济学教授；1988年至1990年间担任普林斯顿大

学经济系主任；1993 年至 1996 年间，先后担任美国总统经济顾问委员会委员，联邦储备委员会副主席。Blinder 发表有多种专著和学术论文，在人力资本和劳动供给、货币政策和财政政策、不确定性和一般均衡等方面都有重要贡献。

Baumol 和 Blinder 合著的《经济学：原理与政策》初版于 1979 年，也是美国等国家最流行的经济学教科书之一（就我所知，本书有澳大利亚版），多次再版，到 1997 年已出到第 7 版。本书的最大特点是叙述上的"问题导向"和"政策导向"。作者认为，学习经济学的最好方法是"从问题到理论"，而不是相反。为此，本书的每一章都从一个具体的、使人感到困惑甚至自相矛盾的政策问题或现实问题开始，然后通过深入浅出的经济学分析，使你走出迷雾。这样做不仅提高了读者读书的兴趣，而且可以加强读者对所学内容的记忆。如果你能记住本书作者所归纳的经济学的 12 个基本观点，你将终身受益。

Paul Samuelson（保罗·萨缪尔森）是当代经济学泰斗，1972 年诺贝尔经济学奖得主，MIT 经济系的创始人。他的《经济学》第 1 版出版于 1948 年，至今已有整整 50 年的历史。在 50 年的时间里，他的《经济学》不断再版，被翻译成包括中文、俄文、日文、法文、德文在内的数十种文字，风靡全世界，是有史以来发行量最大的经济学教科书，教育和影响了数代经济学家、政治家、企业界领袖、新闻记者以及众多的普通人。难怪他自豪地说："如果我能为一个国家写经济学教科书，我并不在乎谁为它写法律，谁在为它起草条约。"Samuelson 本人对他的《经济学》的修订到第 14 版（1992 年）。1995 年，另一位著名经济学家 William Nordhaus 的入盟，使得《经济学》没有随 Samuelson 本人的衰老而衰老，而继续以主流经济学最前沿的教科书身份畅销全球。你现在看到的是该书的第 16 版，出版于第 1 版 50 岁生日之际。

尽管这三本教材所叙述的经济学原理并无大的差别，但由于每个作者的个人偏好和学术风格不同，在叙述方法、切入角度、材料的取舍等方面还是各有千秋。不同书中用于说明同一原理的有些材料互为补充，有助于加深读者对基本理论的理解。据悉，这三本书已由 18 位国际著名商学院的教授联合推荐作为中国 MBA 及经济学/管理学院的首选教科书。我个人学习经济学多年，可以说是经济学的老学生了，但现在再拿起来读这些入门书，仍感有不少新的收获。事实上，当我在为博士生备课时，常常要翻一翻这样一些经济学入门书，以在理解上"温浅而知深"，讲课时深入浅出。美国著名经济学家 Herbert Stein 说：绝大部分可用于指导公共政策的经济学都在本科生的经济学入门教程中。我的经历告诉我，他的话是对的。

如果你的经济条件许可，我建议你不妨同时拥有这三本书。我的经验是，读三本由不同经济学大师写的同一类型的教科书要比读三遍同一大师写的同一本书收获大。你可以从任何一本书开始，然后在每一部分对照阅读。特别地，当你对一本书的内容难以理解时，最好看看另外两本书上是如何写的。你大概有过这样的生活经历：见过一次面的人没有给你留下什么印象，但在另一场合再次谋面后，他就深深印在你的脑海中。读书也是如此。当然，如果经济条件不允许，你买其中的一本就可以了。但你最好能记住我一开始说过的话：**读书最大的成本是时间而不是书价。要买好书，不要买便宜的书。**

现在，你可以作出自己的决策了。

张维迎
北京大学光华管理学院经济学教授
北京大学工商管理研究所所长
英国牛津大学经济学博士
1998 年 6 月 7 日星期日

To my four children, Ellen, Daniel, and now Sabrina and Jim

W.J.B.

To the memory of my father, Morris Blinder, to whom I owe much

A.S.B.

PREFACE

The Seventh Edition continues the basic philosophy of its predecessors. In particular, we avoid the fiction, so popular among textbook writers, that everything is of the utmost importance—a pretense that students are sufficiently intelligent to see through in any event. We try, instead, to highlight those important ideas that are likely to be of lasting significance—principles that students will want to remember long after the course is over because they offer insights that are far from obvious, because they are of practical importance, and because they are widely misunderstood by intelligent laymen. A dozen of the most important of these ideas are selected as **Ideas for Beyond the Final Exam** and are called to the reader's attention when they occur through the use of the symbol .

While all modern economics textbooks abound with "real world" examples, we try to go beyond this by elevating the examples to preeminence. In our view, the policy issue or everyday economic problem ought to lead the student naturally to the economic principle, not the other way around. For this reason, almost every chapter begins with a real policy issue or a practical problem that may seem puzzling or paradoxical to noneconomists. We then proceed to describe the economic analysis required to remove the mystery.

In so doing, we use technical terminology and diagrams only where there is a clear need for them, never for their own sake. Still, economics is a technical subject, and so this is, unavoidably, a book for the desk and not for the bed. We make, however, strenuous efforts to simplify the technical level of the discussion as much as possible without sacrificing content. Fortunately, almost every important idea in economics can be explained in plain English, and this is what we try to do.

Each edition of this book has been influenced by economic events. Reflecting the growing "globalization" of the world economy, the Fifth Edition was substantially revised to exhibit more clearly the international linkages between the U.S. economy and other economies. The Sixth Edition was written just after the collapse of communism. Since the book has long focused on the market mechanism and how it works, it was easy to incorporate these stunning events into the discussion. Both of these features are retained in the Seventh Edition.

But the Seventh Edition is also affected by another critical development—one that cannot be dealt with so definitively: rising inequality in the world's industrialized countries in general, and in the United States in particular. While overall income continues to grow in these countries, more and more of each nation's income and wealth are going to the most affluent portion of the population, while lower-income groups fall further and further behind. This development is not new; it has been going on for more than a decade. But its very persistence has now made it a subject of urgent concern.

Rising inequality has been accompanied by a sharp and continuing slowdown of growth and by a severe stagnation in the rise of American real wages. Unfortunately, no one is quite sure about the causes of these worrisome developments

nor about the means that can be used to ameliorate them. The issues are, therefore, raised at a number of points in this book, and the pertinent facts are reported and discussed. But we are suitably modest in discussing both the underlying causes and the policy issues to which they give rise.

Several new features of this edition are worth noting. The order of the core chapters on microeconomics, Chapters 5 through 8, has been changed in response to suggestions from readers. Their suggestions also led us to add more substantial discussions of consumers' surplus in Chapter 5, price discrimination in Chapter 11, and game theory in Chapter 12. Many more "Policy Debate" boxes, a popular feature of the last edition, have been added to this edition. We have also included a new feature that offers students a somewhat more tangible picture of some of the subjects discussed. New boxes called "You Are There" describe a meeting of the Federal Open Market Committee, an antitrust court case, and activities on the floor of the New York Stock Exchange. Finally, the book has undergone extensive and painstaking updating throughout, so that new material could be added without increasing overall length.

As a last personal note, we must mention that most of the Seventh Edition, like the Sixth before it, had to be written under the handicap of a separation between the coauthors. Happily, this separation entailed neither disagreement nor rancor, and is now over. Alan Blinder's sojourn in Washington—first as a member of the President's Council of Economic Advisers and then as Vice Chairman of the Federal Reserve Board—has undoubtedly contributed to the relevance of this book's materials to the real world. But it has also made communication between us far more challenging.

However, there is a silver lining to this cloud of inconvenience. Blinder reports renewed and enhanced respect for Herbert Stein's sage observation that "most of the economics that is usable for advising on public policy is at about the level of the introductory undergraduate course."

NOTE TO THE STUDENT

We would like to offer one suggestion for success in your economics course. Unlike some of the other courses you may be taking, economics is cumulative—each week's lesson builds upon what you have learned before. You will save yourself both a lot of frustration and a lot of work by keeping up on a week-to-week basis. To assist you in doing so, a chapter summary, a list of important terms and concepts, and a selection of questions to help you with your review are provided at the end of each chapter. Making use of these learning aids will increase your success in your economics course. For additional assistance, see the following list of ancillary materials.

ANCILLARIES

As economic education incorporates new technologies, our extensive learning package has been expanded and improved to accommodate the needs of students and instructors.

- *Study Guide* by Craig Swan, University of Minnesota (for students). Our study guide, which is available in micro and macro split versions, includes learning objectives, a list of important terms and concepts for every chapter, and a quiz that helps students test their understanding and comprehension of concepts. Also included are multiple-choice tests for self-understanding, a list of supplementary readings and study questions for every chapter, and "Economics

in Action" sections that use current news articles to illustrate economic concepts. A new feature, "Economics On-Line," outlines useful Internet and Web sources for economic data and information. Periodic updates of these sources will be provided on the Dryden/Harcourt Web page: www.hbcollege.com.

- *Instructor's Manual/Transparency Masters* by John Isbister, University of California—Santa Cruz. Every chapter in the revised *Instructor's Manual* consists of detailed chapter outlines, teaching tips and suggestions, answers to review questions in the text, and questions for classroom discussion. Multiple-choice questions suitable for quizzes and tests have been added in this edition. The transparency masters include all figures and tables not available as acetates.
- *Microeconomics Test Bank* by Colleen Cameron, University of Mississippi. This updated and reorganized test bank, which consists of more than 5,500 questions, covers Chapters 1–21 and Chapter 34 in the text. Every question has been checked to ensure the accuracy and clarity of the answers, and special thanks go to Bryan Taylor for providing this important service. The questions include true/false, analytical and definitional, multiple-choice, and critical-thinking, short-answer questions. Questions have been organized by section in the text to help instructors pick and choose questions with minimum inconvenience. The test bank is also available in computerized IBM and Windows versions in our easy-to-use *EXAMaster* program. *EXAMaster* allows you to add and edit your own questions, create and edit graphics, print scrambled versions of tests, convert multiple-choice questions to open-ended questions, plus much more.
- *Macroeconomics Test Bank* by John Dodge, University of Sioux Falls. This updated test bank covers Chapters 1–4 and 22–38 in the text. Consisting of more than 5,000 questions, the test bank helps students understand and comprehend the book's concepts and their applicability to real-world situations. Every question has been checked to ensure the accuracy and clarity of the answers, and we thank Ivan Weinel for assisting with this important task. The true/false, multiple-choice, and short-answer questions, which are organized by section in the text, encourage critical thinking and analytical skills. The test bank is also available in computerized IBM and Windows versions in the *EXAMaster* program.
- *Transparency Acetates.* Full color transparency acetates for all important figures and tables in the text are available in micro and macro split versions.
- *Weekly News Updates.* Each news update links topics in the text with current economic events and consists of the synopsis of a pertinent article, references to a topic or chapter in the text, and a few discussion questions suitable for classroom use. The news updates are posted on the Harcourt Web page.
- *PowerPoint® Presentation Software.* This user-friendly slide show is suitable for classroom presentations. It consists of important graphs and tables from the text as well as bulleted summaries and chapter outlines. *PowerPoint Viewer* is provided in the package.
- *Tutorial, Analytical, and Graphing (TAG) Software* by Todd Porter and Teresa Riley, Youngstown State University (for students). In addition to an extensive chapter-by-chapter tutorial, a hands-on graphing section in which students are actually required to draw curves (with keystrokes or a mouse), and a practice exam for each section, this tutorial software consists of a number of new and innovative features added for this edition. Annotated multiple-choice questions give students feedback on their answers and explain reasons

why a particular answer is right or wrong. The software is now available in a new Windows version.

The Dryden Press may provide complimentary instructional aids and supplements or supplement packages to those adopters qualified under its adoption policy. Please contact your sales representative for more information. If as an adopter or potential user you recei.e supplements you do not need, please return them to your sales representative or send them to: Attn: Returns Department, Troy Warehouse, 465 South Lincoln Drive, Troy, MO 63379.

NOTE TO THE INSTRUCTOR

In trying to improve the book from one edition to the next, we rely heavily on our experiences as teachers. But our experience using the book is minuscule compared with that of the hundreds of instructors who use it nationwide. If you encounter problems, or have suggestions for improving the book, we urge you to let us know by writing to either one of us in care of The Dryden Press, 301 Commerce Street, Suite 3700, Fort Worth, TX 76102. Such letters are invaluable, and we are glad to receive them, even if they are critical (but not *too* critical!). Many such suggestions accumulated over the past three years found their way into the Seventh Edition. What follows are suggested course outlines for both one-semester and one-quarter courses.

OUTLINE FOR A ONE-SEMESTER COURSE IN MICROECONOMICS

Chapter Number	Title
1	What Is Economics?
2	The U.S. Economy: Myth and Reality
3	Scarcity and Choice: *The* Economic Problem
4	Supply and Demand: An Initial Look
5	Consumer Choice: The Demand Side of the Market
6	Demand and Elasticity
7	Production, Inputs, and Cost: Building Blocks for Supply Analysis
8	Output, Price, and Profit: The Importance of Marginal Analysis
9	The Firm and the Industry under Perfect Competition
10	The Price System and the Case for Laissez-Faire
11	Monopoly
12	Between Competition and Monopoly
13	The Market Mechanism: Shortcomings and Remedies
14	Real Firms and Their Financing: Stocks and Bonds
15	Pricing the Factors of Production
16	Labor: The Human Input
18	Limiting Market Power: Regulation of Industry
19	Limiting Market Power: Antitrust Policy
20	Taxation and Resource Allocation

Plus any two of the following:

17	Poverty, Inequality, and Discrimination
21	Environmental Protection and Resource Conservation: The Economist's Approach
34	International Trade and Comparative Advantage
38	Comparative Economic Systems: What Are the Choices?

OUTLINE FOR A ONE-SEMESTER COURSE IN MACROECONOMICS

Chapter Number	Title
1	What Is Economics?
2	The U.S. Economy: Myth and Reality
3	Scarcity and Choice: *The* Economic Problem
4	Supply and Demand: An Initial Look
22	The Realm of Macroeconomics
23	Unemployment and Inflation: The Twin Evils of Macroeconomics
24	Income and Spending: The Powerful Consumer
25	Demand-Side Equilibrium: Unemployment or Inflation?
26	Changes on the Demand Side: Multiplier Analysis
27	Supply-Side Equilibrium: Unemployment *and* Inflation?
28	Managing Aggregate Demand: Fiscal Policy
29	Money and the Banking System
30	Monetary Policy and the National Economy
31	The Debate over Monetary Policy
32	Budget Deficits and the National Debt: Fact and Fiction
33	The Trade-off between Inflation and Unemployment
34	International Trade and Comparative Advantage
35	The International Monetary System: Order or Disorder?
36	Macroeconomics in a World Economy
37	Productivity and Growth in the Wealth of Nations
38	Comparative Economic Systems: What Are the Choices?

OUTLINE FOR A ONE-SEMESTER COURSE COVERING BOTH MACROECONOMICS AND MICROECONOMICS

Chapter Number	Title
1	What Is Economics?
2	The U.S. Economy: Myth and Reality
3	Scarcity and Choice: *The* Economic Problem
4	Supply and Demand: An Initial Look
5	Consumer Choice: The Demand Side of the Market
6	Demand and Elasticity
7	Production, Inputs, and Cost: Building Blocks for Supply Analysis
8	Output, Price, and Profit: The Importance of Marginal Analysis
11	Monopoly
12	Between Competition and Monopoly
15	Pricing the Factors of Production
16	Labor: The Human Input
22	The Realm of Macroeconomics
24	Income and Spending: The Powerful Consumer
25	Demand-Side Equilibrium: Unemployment or Inflation?
27	Supply-Side Equilibrium: Unemployment *and* Inflation?

OUTLINE FOR A ONE-QUARTER COURSE IN MICROECONOMICS

Chapter Number	Title
3	Scarcity and Choice: *The* Economic Problem
4	Supply and Demand: An Initial Look

5	Consumer Choice: The Demand Side of the Market
6	Demand and Elasticity
7	Production, Inputs, and Cost: Building Blocks for Supply Analysis
8	Output, Price, and Profit: The Importance of Marginal Analysis
9	The Firm and the Industry under Perfect Competition
10	The Price System and the Case for Laissez-Faire
11	Monopoly
12	Between Competition and Monopoly
13	The Market Mechanism: Shortcomings and Remedies
14	Real Firms and Their Financing: Stocks and Bonds
15	Pricing the Factors of Production
16	Labor: The Human Input
20	Taxation and Resource Allocation

OUTLINE FOR A ONE-QUARTER COURSE IN MACROECONOMICS

Chapter Number	Title
1	What Is Economics?
2	The U.S. Economy: Myth and Reality
3	Scarcity and Choice: *The* Economic Problem
4	Supply and Demand: An Initial Look
22	The Realm of Macroeconomics
23	Unemployment and Inflation: The Twin Evils of Macroeconomics
24	Income and Spending: The Powerful Consumer
25	Demand-Side Equilibrium: Unemployment or Inflation?
26	Changes on the Demand Side: Multiplier Analysis
27	Supply-Side Equilibrium: Unemployment *and* Inflation?
28	Managing Aggregate Demand: Fiscal Policy
29	Money and the Banking System
30	Monetary Policy and the National Economy
31	The Debate over Monetary Policy
32	Budget Deficits and the National Debt: Fact and Fiction
33	The Trade-off between Inflation and Unemployment

OUTLINE FOR A ONE-QUARTER COURSE ON APPLICATIONS OF BOTH MACROECONOMICS AND MICROECONOMICS

Chapter Number	Title
17	Poverty, Inequality, and Discrimination
18	Limiting Market Power: Regulation of Industry
19	Limiting Market Power: Antitrust Policy
20	Taxation and Resource Allocation
21	Environmental Protection and Resource Conservation: The Economist's Approach
23	Unemployment and Inflation: The Twin Evils of Macroeconomics
32	Budget Deficits and the National Debt: Fact and Fiction
34	International Trade and Comparative Advantage
35	The International Monetary System: Order or Disorder?
37	Productivity and Growth in the Wealth of Nations
38	Comparative Economic Systems: What Are the Choices?

WITH THANKS

Finally, and with great pleasure, we turn to the customary acknowledgments of indebtedness. Ours have been accumulating now through seven editions. In these days of specialization, not even a pair of authors can master every subject that an introductory textbook must cover. Our friends and colleagues Albert Ando, Charles Berry, Rebecca Blank, William Branson, the late Lester Chandler, Gregory Chow, Avinash Dixit, Robert Eisner, Susan Feiner, the late Stephen Goldfeld, Claudia Goldin, Ronald Grieson, Daniel Hamermesh, Yuzo Honda, Peter Kenen, Melvin Krauss, Herbert Levine, the late Arthur Lewis, Burton Malkiel, Edwin Mills, Janusz Ordover, Uwe Reinhardt, Harvey Rosen, Laura Tyson, and Martin Weitzman have all given generously of their knowledge in particular areas over the course of seven editions. We have learned much from them, and only wish we had learned more.

Many economists and students at other colleges and universities offered useful suggestions for improvements, many of which we have incorporated into the Seventh Edition. We wish to thank Mordechai Kreinin, Michigan State University; David Bradford, University of New Hampshire; Arthur Diamond, University of Nebraska—Omaha; Norman J. Waitzman, the University of Utah; Philip G. King, San Francisco State University; Thomas G. Watkins, Eastern Kentucky University; James E. Bell, Harris-Stowe State College; Donald N. Baum, University of Nebraska—Omaha; Jim Cox, DeKalb College; David N. Weil, Brown University; Ted W. Chiles, Auburn University at Montgomery; Daniel Vencill, San Francisco State University; and Terence J. Alexander, Iowa State University, for their insightful reviews.

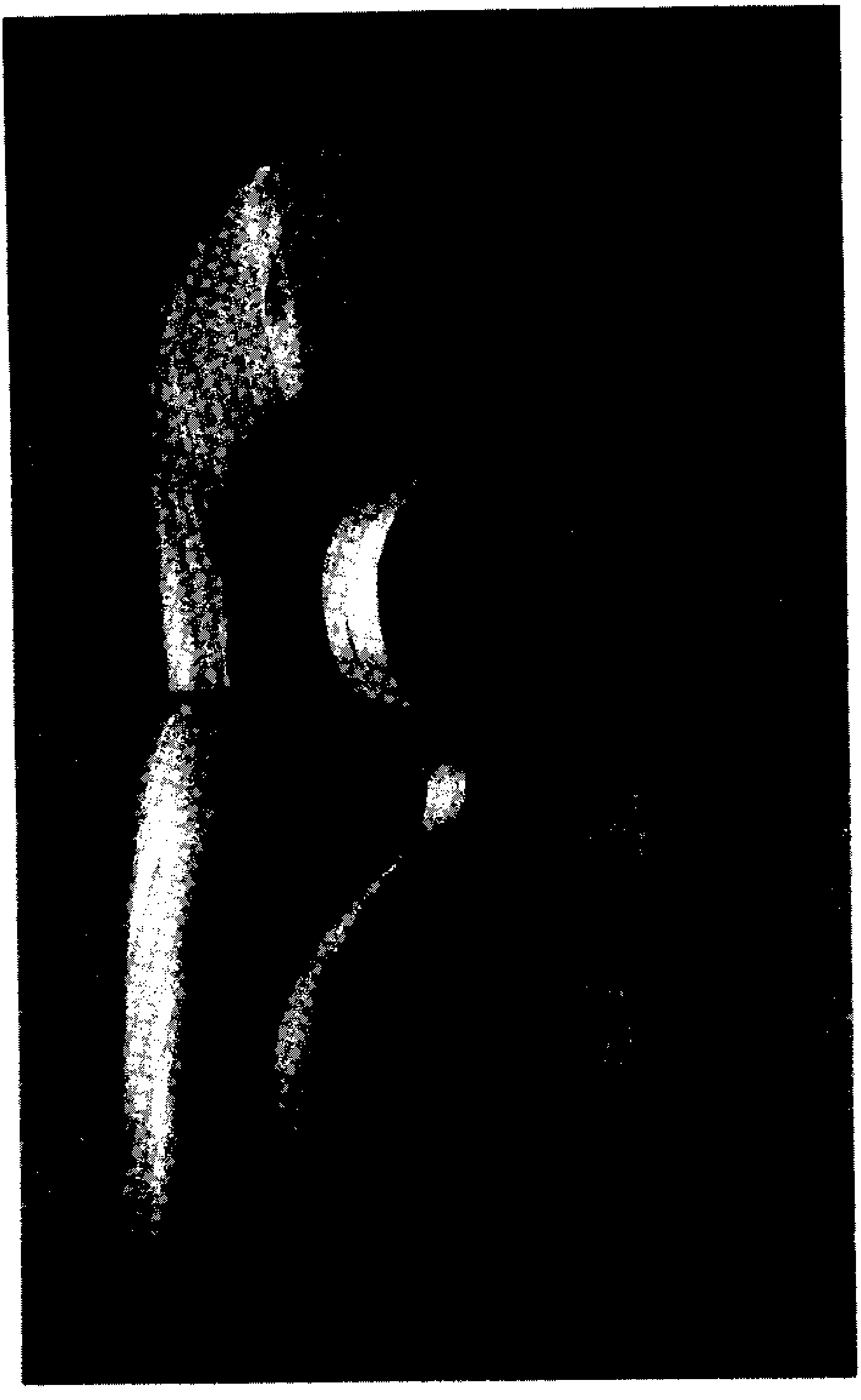
We also wish to thank the many economists who responded to our survey; their responses were invaluable in planning this revision: Donald C. Balch, University of South Carolina; John Bockino, Suffolk Community College; Michael J. Smitka, Washington & Lee University; Hassan Y. Aly, Ohio State University—Marion Campus; Carol M. Clark, Guilford College; John A. Edgren, Eastern Michigan University; Michael Dowd, University of Toledo; Colleen Cameron, University of Southern Mississippi; Donald A. Coffin, Indiana University Northwest; Shyam Bhatia, Indiana University Northwest; John W. Dodge, University of Sioux Falls; John Blair, Wright State University; David Bradford, University of New Hampshire; James H. Breece, University of Maine; Nancy R. Fox, Saint Joseph's University; Terence J. Alexander, Iowa State University; George Giu, Tuskegee University; S. N. Gajanan, University of Pittsburgh at Bradford; Steven E. Abraham, University of Northern Iowa; James N. Wetzel, Virginia Commonwealth University; Bruce Carpenter, Mansfield University; Yilma Gebremariam, Southern Connecticut State University; Marie D. Connolly, Chatham College; and Garry Fleming, Roanoke College.

Obviously, the book you hold in your hand was not produced by us alone. An essential role was played by the fine people at The Dryden Press, including Emily Barrosse, Executive Editor for Economics; Jeanie Anirudhan, Developmental Editor; Stacey Sims, Developmental Editor; Amy Schmidt, Project Editor; Linda Miller, Art Director; and Ann Coburn, Production Manager. We would also like to thank Carol Cirulli and Sean Lanham for their role in this edition. We appreciate all of their efforts. William Baumol is grateful to the publisher, and especially to Linda Miller, the book's art director, for the decision to use photographs of Professor Baumol's sculptures as a design feature.

We also thank our intelligent and delightful secretaries and research coworkers at Princeton University and New York University: Phyllis Durepos and Janeece Roderick, who struggled successfully with the myriad tasks involved in completing the manuscript. Above all, Professor Baumol owes an unrepayable debt to his longstanding partner in crime, Sue Anne Batey Blackman, who carried out much of the updating of materials and who contributed draft paragraphs, illustrative items, and far more with her usual insight and diligence. By now, she undoubtedly knows more about the book than the authors do.

And, finally, there are our wives, Hilda Baumol and Madeline Blinder. They have now participated and helped with this project for over 20 years. During that period, this book has quite literally become part of our families. And both their contributions and our affection have grown.

WILLIAM J. BAUMOL
ALAN S. BLINDER



PART I

Getting Acquainted with Economics

CHAPTER 1

WHAT IS ECONOMICS?

Why does public discussion of economic policy so often show the abysmal ignorance of the participants? Why do I so often want to cry at what public figures, the press, and television commentators say about economic affairs?
Robert M. Solow

Economics is a broad-ranging discipline, both in the questions it asks and the methods it uses to seek answers. Rather than try to define the discipline in a single sentence or paragraph, we will instead introduce you to economics by letting the subject matter speak for itself.

Many of the world's most pressing problems are economic in nature. The first part of the chapter is intended to give you some idea of the sorts of issues that economic analysis helps to clarify and the kinds of solutions that economic principles suggest. The second part briefly introduces the tools that economists use. You are likely to find them useful in your career, personal life, and role as an informed citizen, long after the course is over. Graphs are among these tools, and this book has many of them. For those of you who are unfamiliar with graphs, the appendix to this chapter provides a brief introduction.

IDEAS FOR BEYOND THE FINAL EXAM

As college professors, we realize it is inevitable that you will forget much of what you learn in this course—perhaps with a sense of relief—soon after the final exam. There is not much point bemoaning this fact; elephants may never forget, but people do.

Nevertheless, some economic ideas are so important that you will want to remember them after the course is over. To help you pick out a few of the most crucial concepts, we have selected 12 from among the many contained in this book. Some offer key insights into the workings of the economy. Others bear on important policy issues that appear in newspapers. Others point out common misunderstandings that occur among even the most thoughtful lay observers. As the opening quote of this chapter suggests, many learned judges, politicians, and university administrators who have failed to understand or misused these economic principles could have made wiser decisions than they did.

Each of the **Ideas for Beyond the Final Exam** will be discussed in depth as it occurs in the course of the book, so you should not expect to master them

after reading this opening chapter. Nonetheless, it is useful to sketch them briefly here to introduce you to economics and to provide a preview of what is to come.

IDEA 1 MUTUAL GAINS FROM VOLUNTARY EXCHANGE

One of the most fundamental ideas of economics is that in a *voluntary exchange both* parties must expect to gain something. Otherwise why would they both agree to trade? This principle seems self-evident. Yet it is amazing how often it is ignored in practice.

For example, it was widely believed for centuries that in international trade one country's gain from an exchange must be the other country's loss (Chapter 34). Analogously, some people feel instinctively that if Ms. A profits handsomely from a deal with Mr. B, then Mr. B must have been exploited. Laws sometimes prohibit mutually beneficial exchanges between buyers and sellers—as when a loan transaction is banned because the interest rate is "too high" (Chapter 23), or when a willing worker is condemned to remain unemployed because the wage she is offered is "too low" (Chapter 16), or when the resale of tickets to sporting events ("ticket scalping") is outlawed even though the buyer is happy to get the ticket that he would not obtain at a lower price (Chapter 4).

In every one of these cases, and many more, well-intentioned but misguided reasoning blocks the mutual gains that arise from voluntary exchange—and thereby interferes with one of the most basic functions of an economic system (see Chapter 3).

IDEA 2 RATIONAL CHOICE AND TRUE ECONOMIC COSTS: THE ROLE OF OPPORTUNITY COST

Despite dramatic improvements in our standard of living since the Industrial Revolution, we have not come anywhere near a state of unlimited abundance, and so we must constantly make choices. If you purchase a new home, you may have to give up a trip you had planned. If a firm decides to retool its factories, it may have to postpone plans for new executive offices. If a government expands its defense program, it may be forced to reduce its outlays on roads or school buildings.

Economists say that the true costs of such decisions are not the number of dollars spent on the house, the new equipment, or the military establishment, but rather *the value of what must be given up in order to acquire the item*—the vacation trip, the new executive offices, the improved roads, and the new schools. These are called **opportunity** costs because they represent the *opportunities* the individual, firm, or government must forgo to make the desired expenditure. Economists maintain that rational decision making requires that opportunity costs be considered as part of the cost of any decision (Chapter 3).

The **OPPORTUNITY COST** of some decision is the value of the next best alternative which you have to give up because of that decision (for example, working instead of going to school).

The cost of a college education provides a vivid example that is probably close to your heart. How much do you think it *costs* to go to college? Most likely you would answer this question by adding together your expenditures on tuition, room and board, books, and the like, and then deducting any scholarship funds you may receive. Economists would not. They would first want to know how

much you could be earning if you were not attending college. This may sound like an irrelevant question; but because you give up these earnings by attending college, they must be added to your tuition bill as a cost of your education. Nor would economists accept the university's bill for room and board as a measure of your living costs. They would want to know by how much this exceeds what it would have cost you to live at home, and only this extra cost would be counted as an expense. On balance, a college education probably costs more than you think.

IDEA 3 ATTEMPTS TO REPEAL THE LAWS OF SUPPLY AND DEMAND: THE MARKET STRIKES BACK

When a commodity is in short supply, its price naturally tends to rise. Sometimes disgruntled consumers badger politicians into "solving" the problem by imposing a legal ceiling on the price. Similarly, when supplies are abundant—say, when fine weather produces extraordinarily abundant crops—prices tend to fall. This naturally dismays producers, who often succeed in getting legislation to prohibit low prices by imposing price floors.

But such attempts to repeal the laws of supply and demand usually backfire, and sometimes produce results virtually the opposite of those that were intended. Where rent controls are adopted to protect tenants, housing grows scarce because the law makes it unprofitable to build and maintain apartments. When price floors are placed under agricultural products, surpluses pile up.

As we will see in Chapter 4 and elsewhere in this book, such consequences of interfering with the price mechanism are no accident. They follow inevitably from the way free markets work.

IDEA 4 THE IMPORTANCE OF MARGINAL ANALYSIS

We will devote many pages of this book to explaining, and extolling the virtues of, a type of decision-making process called *marginal analysis* (see especially Chapters 7 and 8), which can best be illustrated by an example.

Suppose an airline is told by its accountants that the full cost of transporting one passenger from Los Angeles to New York is \$300. Can the airline profit by offering a reduced fare of \$200 to students who fly on a standby basis? The surprising answer is: probably yes. And the reason is that most of the costs will be paid whether the plane carries 20 passengers or 120 passengers.

Marginal analysis points out that costs such as maintenance, landing rights, and ground crews are irrelevant to the decision about whether to carry standby passengers for reduced rates. The only costs that *are* relevant are the *extra* costs of writing and processing additional tickets, the food and beverages these passengers consume, the additional fuel required, and so on. These costs are called *marginal costs* and are probably quite small in this example. Any passenger who pays the airline more than its marginal cost will add something to the company's profit. So it probably is more profitable to let the students ride at low fares than to let the plane fly with empty seats.

There are many real cases in which decision makers, not understanding marginal analysis, have rejected such advantageous possibilities as the reduced fare

in our example. These people were misled by calculating in terms of *average* rather than *marginal* cost figures—an error that can be quite costly.

IDEA 5 EXTERNALITIES: A SHORTCOMING OF THE MARKET CURED BY MARKET METHODS

Markets are very efficient at producing just the goods that consumers want and in the quantities they desire. They do so by rewarding those who respond to what consumers want and who produce these products economically. Similarly, the market mechanism ferrets out waste and inefficiency by seeing to it that inefficient producers lose money.

This works well as long as an exchange between a seller and a buyer affects only those two parties. But often an economic transaction affects uninvolved third parties. Examples abound. The electric utility that generates power for midwestern states also produces pollution which kills freshwater fish in New York state. A farmer sprays crops with toxic pesticides, but the poison seeps into the ground water and affects the health of neighboring communities.

Such social costs—called *externalities* because they affect parties *external* to the economic transactions that cause them—escape the control of the market mechanism. As we will learn in Chapters 13 and 21, there is no financial incentive to motivate polluters to minimize the damage they do. Hence, business firms make their products as cheaply as possible, disregarding externalities that may damage the quality of life.

Yet Chapters 13 and 21 point out a way for the government to use the market mechanism to control undesirable externalities. If the electric utility and the farmer are charged for the harm they cause the public, just as they are charged when they use tangible resources such as coal and fertilizer, then they will have an incentive to reduce the amount of pollution they generate. Thus, in this case, economists believe that market methods are often the best way to cure one of the market's most important shortcomings.

IDEA 6 THE COST DISEASE OF THE PERSONAL SERVICES

A distressing phenomenon is occurring throughout the industrialized world. Many community services have been deteriorating—growing numbers of people find it harder to pay for health care, there are fewer postal deliveries and garbage pickups and larger classes in public schools—even though the public is paying more for them. The costs of providing many services have risen consistently faster than the rate of inflation.

Perhaps the most publicized examples are medical care and education. For example, over a 47-year period the daily cost of a stay at a hospital outstripped the rate of inflation by 800 percent! A natural response is to attribute the problem to greed, inefficiency, and political corruption. But this cannot be the whole story, because the scenario has been repeated in virtually every other industrialized country, in spite of the great differences in the way health-care services are provided.

As we shall see in Chapter 13, one of the major causes of the problem is economic. And it stems from the dazzling growth in efficiency of private manufacturing industries. Because technological improvements make workers more productive in manufacturing, costs go down and wages rise. But wages rise not only

for the manufacturing workers but also for hospital workers, teachers, and other service workers (because otherwise these workers would leave their low-paying service jobs and compete for jobs in high-paying industries). Unlike manufacturing, however, the technology of labor-intensive personal services is not easily changed. Since it still takes one person to drive a postal truck and one teacher to teach a class, the cost of these services is forced to rise in step with wage increases. This is what has been called the "cost disease" of the personal services, a malady that affects many services, including university teaching, restaurant cooking, retailing, and automobile repair.

This is important to understand not because it excuses the financial record of our governments, but because it suggests what we should expect the future to bring and, perhaps, indicates what policies we should advocate to deal with it.

IDEA 7 THE TRADE-OFF BETWEEN OUTPUT AND EQUALITY

Several much-publicized studies in 1995 confirmed that inequality in income and wealth has grown substantially for several decades, so that the rich have steadily grown richer while the poor have become poorer. Moreover, the studies indicated that inequality was more extreme in the United States than in any of the world's other wealthy countries. At the same time, Congress was undertaking changes in the federal tax system that were designed to spur productivity and efficiency by providing greater incentives for business activity. Often, that means lowering tax rates for the rich. There are important elements of truth in this position. No one doubts that proper economic incentives are critical. Yet many people feel that the unequal distribution of income in our society is unjust.

In fact, we have a genuine dilemma. To provide stronger incentives for success in the economic game, the gaps between the "winners" and the "losers" may have to be widened. It is these gaps, after all, that provide the incentives to work harder, to save more, and to invest productively. But such programs also breed inequality. Thus, economists say there is a *trade-off* between the *size* of a nation's output and the degree of *equality* with which that output is distributed. Supply-side tax cuts are one example. Another is antipoverty programs. As we will see in Chapter 17, many policies designed to divide the proverbial economic pie more equally inadvertently can cause the size of the pie to shrink.

IDEA 8 THE TRADE-OFF BETWEEN INFLATION AND UNEMPLOYMENT

Inflation these days is running a bit below 3 percent per year. In 1990, it was just over 5 percent. What made the inflation rate fall? Most economists believe the answer is simple: In mid-1990, the U.S. economy entered a recession that proved to be deeper and longer than policymakers anticipated. The unemployment rate rose from a low of 5 percent in mid-1990 to a peak of nearly 8 percent in mid-1992 and then receded only slowly.

Economists maintain that this conjunction of events—falling inflation and high unemployment—was no coincidence. Owing to features of our economy that we will study in Parts VI and VIII, there is an agonizing *trade-off between inflation and unemployment*, meaning that most policies that lower inflation also cause higher unemployment for a while.

Since this trade-off poses one of the fundamental dilemmas of national economic policy, we will devote all of Chapter 33 to examining it in detail. And we

shall also consider some suggestions for escaping from the trade-off, such as supply-side economics (Chapter 27).

IDEA 9 THE ILLUSION OF HIGH INTEREST RATES

Is it more costly to borrow money at 12 percent interest or at 8 percent? That question seems easy, even without a course in economics. But, in fact, it is not. An example will show why.

In 1995, banks were lending money to home buyers at annual interest rates below 8 percent. In 1980, these rates had been over 12 percent. Yet economists maintain that it was actually cheaper to borrow in 1980 than in 1995. Why? Because inflation in 1980 was running at about 10 percent per year while it was just below 3 percent by 1995.

But why is the rate of inflation relevant for deciding how costly it is to borrow? Consider the position of a person who borrows $100 for 1 year at a 12 percent rate of interest when prices are rising at 10 percent per year. At the end of the year the borrower pays back her $100 plus $12 interest. But over that same year her indebtedness declines by $10 *in terms of what that money will buy.* Because the purchasing power of money declines 10 percent, it is as if $10 of her debt had been forgiven. Thus, in terms of *purchasing power,* the borrower really pays only $2 in interest on her $100 loan, or 2 percent.

Now consider someone who borrows $100 at 8 percent interest when inflation is only 3 percent. This borrower pays back the original $100 plus $8 in interest and sees the purchasing power of his debt decline by $3 due to inflation—for a net payment in purchasing-power terms of $5, or 5 percent. Thus, in the economically relevant sense, the 8 percent loan at 3 percent inflation is actually more expensive than the 12 percent loan at 10 percent inflation.

As we will learn in Chapter 23, the failure to understand this principle has caused problems for our tax laws, for the financial system, and for public utilities. In Chapter 32 we will see that it has even led to misunderstanding of the size and nature of the government budget deficit.

IDEA 10 DO BUDGET DEFICITS BURDEN FUTURE GENERATIONS?

The political struggle to reduce the federal budget deficit has been in the news for 15 years now. First President Reagan and then President Bush argued that raising taxes is worse than tolerating the deficit, a claim disputed by many. Critics objected that deficits hold dire consequences—including higher interest rates, more inflation, a stagnant economy, and an irksome burden on future Americans. Finally, during the Clinton administration, a large deficit-reduction package was enacted in 1993, and a political brouhaha arose in 1995–1996 over how to balance the budget.

The conflicting claims and counterclaims that have marked this debate are bound to confuse the layperson. Who is right? Are deficits really malign or benign influences on our economy? The answers, economists insist, are so complicated that the only correct short answer is: It all depends. The precise factors on which the answers depend, and the reasons why, are important enough to merit an entire chapter (Chapter 32). There we will learn that a budget deficit may or may not burden future generations, depending on its size and on the reasons for its existence.

IDEA 11 THE OVERWHELMING IMPORTANCE OF PRODUCTIVITY GROWTH IN THE LONG RUN

In Geneva, a worker in a watch factory now turns out roughly 100 times as many mechanical watches per year as her ancestors did three centuries earlier. The *productivity* of labor (output per hour of work) in cotton production has probably gone up more than 1,000-fold in 200 years. It is estimated that rising labor productivity has increased the standard of living of a typical American worker about 7-fold in the past century. This means that Americans now enjoy about seven times as much clothing, housewares, and luxury goods as did a typical inhabitant of the United States 100 years ago.

Economic issues such as inflation, unemployment, and monopoly are important to us all and will receive much attention in this book. But in the long run nothing has as great an effect on our material well-being and the amounts society can afford to spend on hospitals, schools, and social amenities as the rate of growth of productivity. Chapter 37 points out that what appears to be a small increase in productivity growth can have a huge effect on a country's standard of living over a long period of time because productivity compounds like the interest on savings in a bank. Similarly, a slowdown in productivity growth that persists for a substantial number of years can have a devastating effect on standards of living.

IDEA 12 THE SURPRISING PRINCIPLE OF COMPARATIVE ADVANTAGE

The Japanese economy produces many products that Americans buy in huge quantities—including cars, TV sets, cameras, and electronic equipment. American manufacturers often complain about the competition and demand protection from the flood of imports that, in their view, threatens American standards of living. Is this view justified?

Economists think not. They maintain that both sides must gain from international trade. But what if the Japanese were able to produce *everything* more cheaply than we can? Would it not then be true that Americans would be thrown out of work and that our nation would be impoverished?

A remarkable result, called the law of *comparative advantage,* shows that even in this extreme case, the two nations can still benefit by trading and that each can gain as a result! We will explain this principle fully in Chapter 34, where we will also note some potentially valid arguments in favor of providing special incentives for particular domestic industries. But for now a simple parable will make the reason clear.

Suppose Sally grows up on a farm and is a whiz at plowing, but she is also a successful country singer who earns $4,000 a performance. Should Sally turn down singing engagements to leave time for plowing? Of course not. Instead she should hire Alfie, a much less efficient farmer, to do the plowing for her. Sally may be a better farmer. But she earns so much more by specializing in singing that it makes sense to leave the farming to Alfie. Alfie, though a less skilled farmer than Sally, is an even worse singer. Thus, Alfie earns a living by specializing in the job at which he at least has a *comparative* advantage (his farming is not as bad as his singing), and both Alfie and Sally gain. The same is true of two countries. Even if one of them is more efficient at everything, both countries can gain by producing the things they do best *comparatively.*

EPILOGUE

These, then, are a dozen of the more fundamental concepts you will find in this book—ideas that we hope you will retain **Beyond the Final Exam.** There is no need to master them right now, for you will hear much more about each as the book progresses. Instead, keep them in mind as you read—we will point them out to you as they occur—and look back over this list at the end of the course. You may be amazed to see how natural, or even obvious, they will seem then.

INSIDE THE ECONOMIST'S TOOL KIT

Now that you have some idea of the kinds of issues economists deal with, you should know something about the way they grapple with these problems.

ECONOMICS AS A DISCIPLINE

Economics has something of a split personality. Although clearly the most rigorous of the social sciences, it nevertheless looks decidedly more "social" than "scientific" when compared with, say, physics. An economist must be a jack of several trades, borrowing modes of investigation from numerous fields.

Mathematical reasoning is used extensively in economics, but so is historical study. And neither looks quite the same as when practiced by a mathematician or a historian. Statistical inference plays a major role in modern economic inquiry; but economists have had to modify standard statistical procedures to fit the kinds of data they analyze.

An introductory course in economics will not make you an economist; but it should help you approach social problems dispassionately. You will not find solutions to all society's economic problems in this book. But you should learn how to pose questions in ways that will help produce answers that are both useful and illuminating.

THE NEED FOR ABSTRACTION

Some students find economics unduly abstract and "unrealistic." The stylized world envisioned by economic theory seems only a distant cousin to the world they know. There is an old joke about three people—a chemist, a physicist, and an economist—stranded on an island with an ample supply of canned food but no tools to open the cans. The chemist thinks that lighting a fire under the cans would expand their contents and cause the cans to burst. The physicist advocates building a catapult with which they could smash the cans against some nearby boulders. The economist's suggestion? "Assume we have a can opener."

Economic theory *does* make unrealistic assumptions; you will encounter many of them in the pages that follow. But this propensity to abstract from reality results from the incredible complexity of the economic world, not from any fondness economists have for sounding absurd.

Compare the chemist's simple task of explaining the interactions of compounds in a chemical reaction with the economist's complex task of explaining

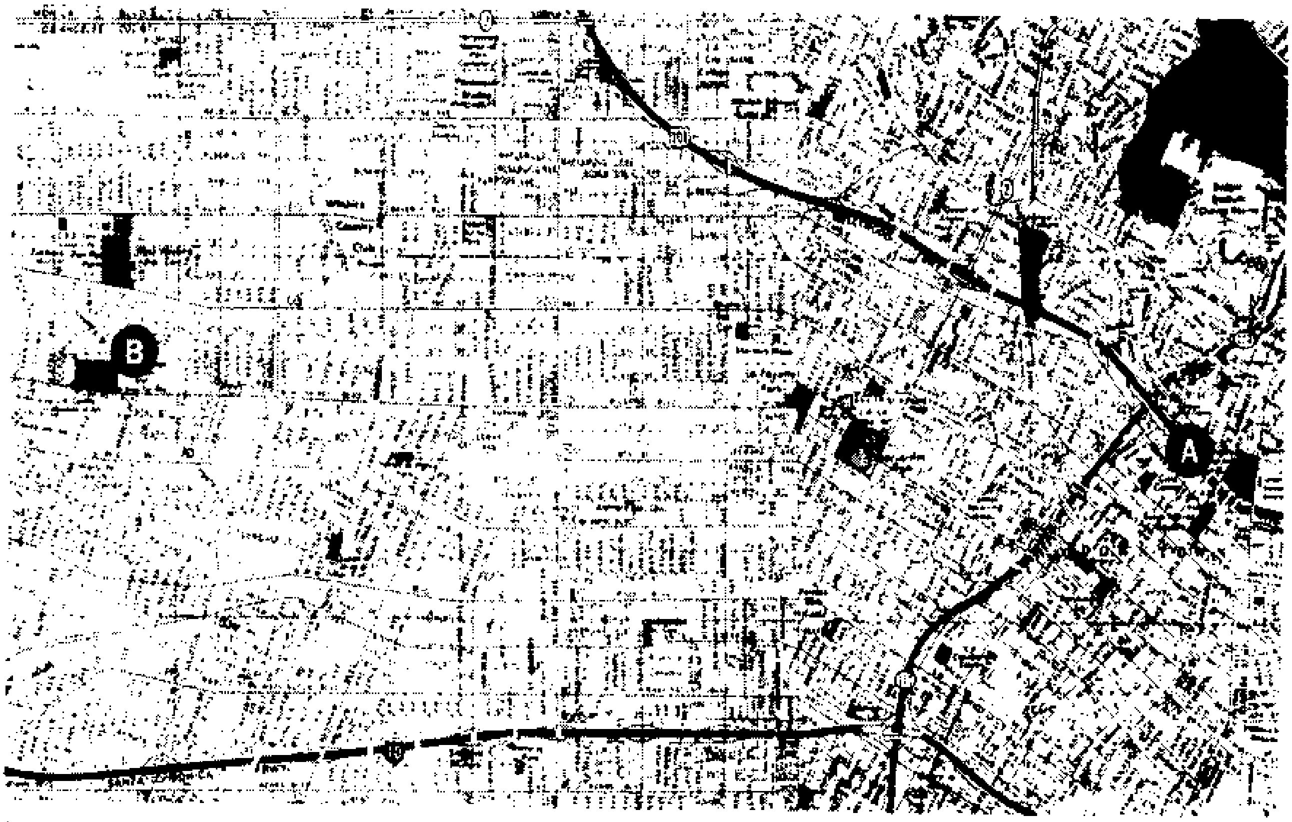
the interactions of people in an economy. Are molecules motivated by greed or altruism, by envy or ambition? Do they ever emulate other molecules? Do forecasts about them influence their behavior? People, of course, do all these things, and many, many more. It is therefore vastly more difficult to predict human behavior than to predict chemical reactions. If economists tried to keep track of every aspect of human behavior, they would never get anywhere. Thus:

Abstraction from unimportant details is necessary to understand the functioning of anything as complex as the economy.

ABSTRACTION means ignoring many details in order to focus on the most important elements of a problem.

To appreciate why economists **abstract** from details, imagine the following hypothetical situation. You have just arrived, for the first time in your life, in Los Angeles. You are now at the Los Angeles Civic Center. This is the point marked *A* in Figures 1-1 and 1-2, which are alternative maps of part of Los Angeles. You want to drive to the Los Angeles County Museum of Art, marked *B* on each map. Which map would you find more useful? You will notice that Map 1 (Figure 1-1) has the full details of the Los Angeles road system. Consequently, it requires a major effort to read it. In contrast, Map 2 (Figure 1-2) omits many minor roads so that the freeways and major arteries stand out more clearly.

FIGURE 1-1 MAP 1



Map 1 gives complete details of the road system of Los Angeles. If you are like most people, you will find it hard to read and not very useful for figuring out how to get from the Civic Center (point *A*) to the Los Angeles County Museum of Art (point *B*). For this purpose, the map is far too detailed, though for some other purposes (for example, locating some small street in Hollywood), it may be the best map available.

SOURCE: © MCMLXXV by North American Maps, P.O. Box 5850, San Francisco, CA 94101.

Most strangers to the city would prefer Map 2. With its guidance, they are likely to find the museum in a reasonable amount of time, even though they might have found a slightly shorter route by careful calculation and planning using Map 1. Map 2 seems to *abstract* successfully from a lot of confusing details while retaining the essential aspects of the city's geography. Economic theories strive to do the same thing.

Map 3 (Figure 1-3), which shows little more than the major interstate routes that pass through the greater Los Angeles area, illustrates a danger of which all theorists must beware. Armed only with the information provided on this map, you might never find your way to the art museum. Instead of a useful idealization of the Los Angeles road network, the map makers have produced a map that is oversimplified for our purpose. Too much has been assumed away. Of course, this map was never intended to be used as a detailed tourist guide, which brings us to an important point:

There is no such thing as one "right" degree of abstraction for all analytic purposes. The proper degree of abstraction depends on the objective of the analysis. A model that is a gross oversimplification for one purpose may be needlessly complicated for another.

Economists are constantly treading the thin line between Map 2 and Map 3, between useful generalization about complex issues and gross distortions of the pertinent facts. How can they tell when they have abstracted from reality just

FIGURE 1-2

MAP 2

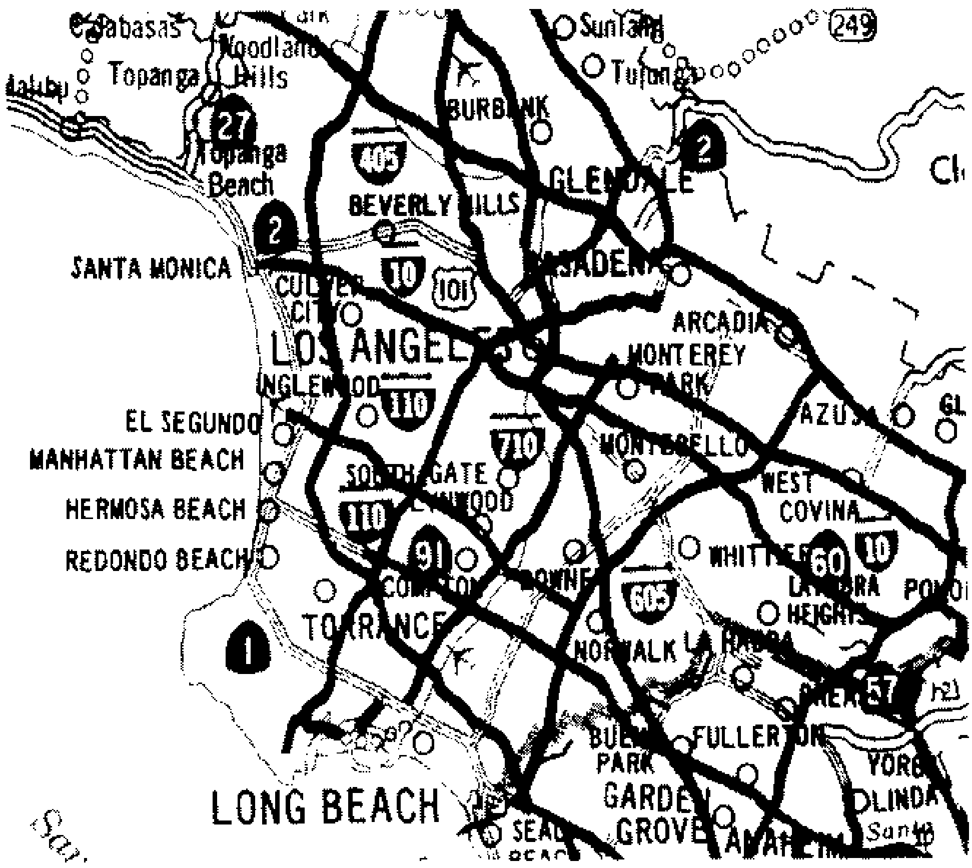


Map 2 shows a different perspective of Los Angeles. Minor roads are eliminated—we might say, *assumed away*—in order to present a clearer picture of where the major arteries and freeways go. As a result of this simplification, several ways of getting from the Civic Center (point *A*) to the Los Angeles County Museum of Art (point *B*) stand out clearly. For example, we can take the Hollywood Freeway west to Alvarado Blvd. south, then west on Wilshire Blvd. The museum is on the right. While we might find a shorter route by poring over the details of Map 1, most of us will feel more comfortable with Map 2.

FIGURE 1-3

Map 3 strips away still more details of the Los Angeles road system. In fact, only major trunk roads and freeways remain. This map may be useful for passing through the city or getting around it, but it will not be of much help to the tourist who wants to see the sights of Los Angeles. For this purpose, too many details are missing.

MAP 3



enough? There is no objective answer to this question, which is why applied economics is as much art as science. One of the things distinguishing good economics from bad economics is the degree to which analysts are able to find the factors that constitute the equivalent of Map 2 (rather than Maps 1 or 3) for the problem at hand. It is not always easy to do.

For example, suppose you want to learn why different people have different incomes, why some are fabulously rich while others are abjectly poor. People differ in many ways, too many to enumerate, much less to study. The economist must ignore most of these details in order to focus on the important ones. The color of a person's hair or eyes is probably not important to the problem at hand, but the color of his or her skin certainly is. Height and weight may not matter, but education probably does. Proceeding in this way, we pare Map 1 down to the manageable dimensions of Map 2. But there is a danger of going too far, stripping away some of the crucial factors, and winding up with Map 3.

THE ROLE OF ECONOMIC THEORY

A person "can stare stupidly at phenomena; but in the absence of imagination they will not connect themselves together in any rational way." These words of the renowned American philosopher-scientist C. S. Peirce succinctly express the crucial role of theory in scientific inquiry. What do we mean by a theory?

A **THEORY** is a deliberate simplification of relationships whose purpose is to explain how those relationships work.

To an economist or natural scientist, the word **theory** means something different from what it means in common parlance. In scientific usage, a theory is *not* an untested assertion of alleged fact. The statement that vitamin E provides protection against cancer is not a theory; it is a *hypothesis*, which will prove to be true or false once the right sorts of experiments have been completed. Instead, a theory is a deliberate simplification (abstraction) of factual relationships that attempts to explain how those relationships work. It is an *explanation* of the mechanism behind observed phenomena. Thus, gravity forms the basis of theories that describe and explain the paths of the planets. Similarly, Keynesian theory (discussed in Parts VI and VIII) seeks to describe and explain how government policies affect the path of the national economy.

People who have never studied economics often draw a false distinction between *theory* and *practical policy*. Politicians and business people, in particular, often reject abstract economic theory as something that is best ignored by "practical" people. The irony of these statements is that:

It is precisely the concern for policy that makes economic theory so necessary and important.

If we could not change the economy through public policy, economics could be a historical and descriptive discipline, asking, for example, what happened in the United States during the Great Depression of the 1930s or how is it that industrial pollution got to be so serious in the 20th century. But deep concern about public policy forces economists to go beyond historical questions. To analyze policy options, they are forced to deal with possibilities *that have not actually occurred.*

For example, to learn how to prevent economic depressions, they must investigate whether the Great Depression of the 1930s could have been avoided by more astute government policies. Or to determine what environmental programs will be most effective, they must understand how and why a market economy produces pollution and what might happen if government placed taxes on industrial waste discharges and automobile emissions.

Two variables are said to be **CORRELATED** if they tend to go up or down together. But correlation need not imply causation.

The facts, however, can sometimes be highly misleading. Data often indicate that two variables move up and down together. But this statistical **correlation** does not prove that either variable *causes* the other. For example, people drive their cars more slowly when it rains, and there are also more traffic accidents. But this correlation does not mean that slow driving causes accidents. Rather, we understand that both phenomena are caused by a common underlying factor—more rain. How do we know this? Not just by looking at the correlation (the degree of similarity) between data on accidents and driving speeds. Data alone tell us little about cause and effect. We must use some simple theory as part of our analysis.

Similarly, most economic issues hinge on some question of cause and effect. So simply observing correlations in data is not enough. Only a combination of theoretical reasoning and data analysis can hope to provide useful answers. We must first proceed deductively from assumptions to conclusions and then test the conclusions against data. In that way, we may hope to understand *how*, if at all, different government policies will lead to a lower unemployment rate or *how* a tax on emissions will reduce pollution.

Statistical correlation need not imply causation. Some theory is usually needed to interpret data.

WHAT IS AN ECONOMIC "MODEL"?

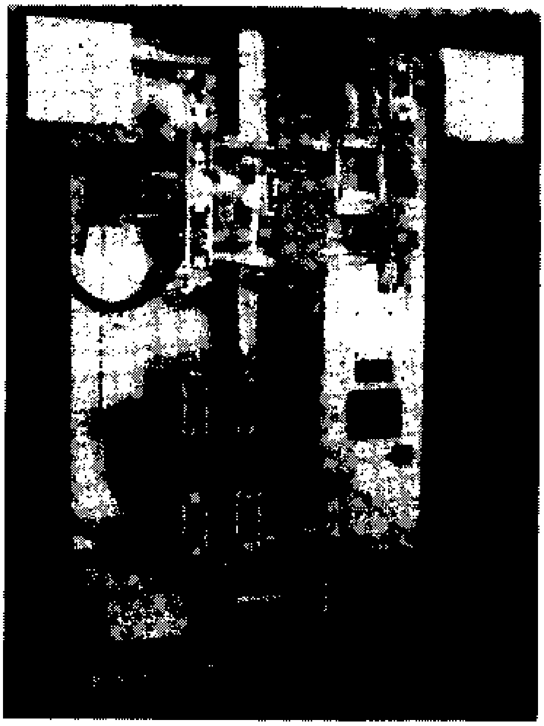
An **ECONOMIC MODEL** is a simplified, small-scale version of some aspect of the economy. Economic models are often expressed in equations, by graphs, or in words.

An economic model is a representation of a theory or a part of a theory, often used to gain insight into cause and effect. The notion of a "model" is familiar enough to children; and economists—like other scientists—use the term in much the same way that children do.

A child's model automobile or airplane looks and operates much like the real thing, but it is much smaller and much simpler, and so it is easier to manipulate and understand. Engineers for General Motors and Boeing also build models of cars and planes. While their models are far bigger and much more elaborate than a child's toy, they use them for much the same purposes: to observe the workings of these vehicles "up close," to experiment with them in order to see how they might behave under different circumstances. ("What happens if I do this?") From these experiments, they make educated guesses as to how the real-life version will perform.

Economists use models for similar purposes. A. W. Phillips, the famous engineer-turned-economist who discovered the "Phillips curve" (discussed in Chapter 33), was talented enough to construct a working model of the determination of national income in a simple economy, using colored water flowing through pipes. For years this contraption, depicted in Figure 1-4, has graced the basement of the London School of Economics. However, most economists lack Phillips's manual dexterity, so economic models are generally built with paper and pencil rather than with hammer and nails.

FIGURE 1-4 THE PHILLIPS MACHINE



The late Professor A. W. Phillips, while teaching at the London School of Economics in the early 1950s, built this machine to illustrate Keynesian theory. This is the same theory that we will explain with words and diagrams later in the book; but Phillips's background as an engineer enabled him to depict the theory with the help of tubes, valves, and pumps. Because economists are not very good plumbers, few of them try to build models of this sort; most rely on paper and pencil instead. But the two sorts of models fulfill precisely the same role. They simplify reality in order to make it understandable.

Because many of the models used in this book are depicted in diagrams, we explain the construction and use of various types of graphs in the appendix to this chapter. But sometimes economic models are expressed only in words. The statement "Business firms produce the level of output that maximizes their profits," is the basis for a behavioral model whose consequences we explore in some detail in Parts II through V. Don't be put off by seemingly abstract models. Think of them as useful road maps. And remember how hard it would be to find your way around Los Angeles without one.

REASONS FOR DISAGREEMENTS: IMPERFECT INFORMATION AND VALUE JUDGMENTS

"If all the earth's economists were laid end to end, they could not reach an agreement," or so the saying goes. Politicians and reporters are fond of pointing out that economists can be found on both sides of many public policy issues. If economics is a science, why do economists quarrel so much? After all, astronomers do not debate whether the earth revolves around the sun or vice versa.

The question reflects a misunderstanding of the nature of science. Disputes are normal at the frontier of any science. For example, astronomers once did argue, and quite vociferously, over whether the earth revolves around the sun. Nowadays, they argue about black holes, gamma-ray bursts, brown dwarfs, and other esoterica. These arguments go mostly unnoticed by the public because few of us understand what they are talking about. But economics is a *social* science, so its disputes are aired in public. All sorts of people are eager to join economic debates about inflation, pollution, poverty, and the like.

Furthermore, the fact is that economists agree on much more than is commonly supposed. Virtually all economists, regardless of their politics, agree that taxing polluters is one of the best ways to protect the environment (see Chapters 13 and 21), that rent controls can ruin a city (Chapter 4), and that free trade among nations is preferable to the erection of barriers through tariffs and quotas (see Chapter 34). The list could go on and on. It is probably true that the issues about which economists agree *far* exceed the subjects on which they disagree.

Finally, many disputes among economists are not scientific disputes at all. Sometimes the pertinent facts are simply unknown. For example, you will learn in Chapter 21 that the proper tax to levy on industrial wastes depends on quantitative estimates of the harm done by the pollutant. Unfortunately, for most waste products, good estimates are not yet available. This makes it difficult to agree on a concrete policy proposal.

Another important source of disagreements is that economists, like other people, come in all political stripes: conservative, middle-of-the-road, liberal, radical. Each may have different values and a different view of what constitutes a good society. So each may hold a different view of the "right" solution to a public policy problem, even if they agree on the underlying analysis.

For example, we noted early in this chapter that anti-inflation policies are likely to cause recessions. Using tools we will describe in Part VII, many economists believe they can even measure how deep a recession we must endure to reduce inflation by a given amount. Is it worth having 2.6 million more people out of work for a year to cut the inflation rate by 1 percent? An economist cannot answer this any more than a nuclear physicist could have determined whether dropping the atomic bomb on Hiroshima was a good idea. The decision rests on judgments about the moral trade-off between inflation and unemployment, judgments that can be made only by the citizenry through its elected officials.

While economic science can contribute the best theoretical and factual knowledge there is on a particular issue, the final decision on policy questions often rests either on information that is not currently available or on tastes and ethical opinions about which people differ (the things we call "value judgments"), or on both.

LAST WORD: COMMON SENSE IS NOT ALWAYS RELIABLE

Many people think sound decisions are just a matter of "common sense." If that were so—if untrained but intelligent observers could reach the right economic decisions using only their instincts and intuition—there would be little reason to study economics. Unfortunately, common sense is not always a reliable guide in economics.

True, there are many cases where it is not misleading. Most people undoubtedly realize, for example, that a surge in demand for a product is likely to raise its price, at least for a while. They also understand that increases in the prices of American goods will reduce the quantity we can export to foreign countries.

But many economic relationships are counterintuitive. Try your intuition on this one, for example. You own a widget manufacturing company that rents a warehouse. Your landlord raises your rent by $10,000 per year. Should you raise the price of your widgets to try to recoup some of your higher costs? Or should you lower your price to try to sell more and "spread your overhead?" We shall see in Chapter 7 that both answers are probably wrong!

We will explain many more counterintuitive economic relationships in this book. By the end, you will have a better sense of when common sense works and when it fails. You will be able to recognize common fallacies that are all too often offered as pearls of wisdom by public figures, the press, and television commentators.

SUMMARY

1. To help you get the most out of your first course in economics, we have devised a list of 12 important ideas that you will want to remember **Beyond the Final Exam.**
2. Economics is a broad-ranging discipline that uses a variety of techniques and approaches to address important social questions.
3. Because of the great complexity of human behavior, economists are forced to **abstract** from many details, to make *generalizations* that they know are not quite true, and to organize what knowledge they have in terms of some theoretical structure called a "model."
4. **Correlation** need not imply *causation.*
5. Economists use simplified models to understand the real world and predict its behavior, much as a child uses a **model** railroad to learn how trains work.
6. While these models, if skillfully constructed, can illuminate important economic problems, they rarely can answer the questions that confront policymakers. Value judgments are needed for this purpose, and the economist is no better equipped than anyone else to make them.
7. Common sense is often an unreliable guide to the right economic decision.

KEY TERMS

Voluntary exchange
Opportunity cost
Comparative advantage
Productivity
Externalities
Marginal analysis
Marginal costs
Abstraction and generalization
Theory
Correlation versus causation
Economic model

QUESTIONS FOR REVIEW

1. Think about how you would construct a "model" of how your college is governed. Which officers and administrators would you include and exclude from your model if the objective were:
   a. to explain how decisions on financial aid are made?
   b. to explain the quality of the faculty?
   Relate this to the map example in the chapter.
2. Relate the process of "abstraction" to the way you take notes in a lecture. Why do you not try to transcribe every word the lecturer utters? Why do you not just write down the title of the lecture and stop there? How do you decide, roughly speaking, on the correct amount of detail?
3. Explain why a government policymaker cannot afford to ignore economic theory.

APPENDIX THE GRAPHS USED IN ECONOMIC ANALYSIS[1]

Economic models are frequently analyzed and explained with the help of graphs; and this book is full of graphs. But that is not the only reason for you to study how they work. Most of you will deal with graphs in the future, perhaps frequently. They appear in newspapers. Governments use them to keep track of the amount of money that they owe to foreign countries. Business firms use them to check their profit and sales performance. Persons concerned with social issues use them to examine trends in ethnic composition of cities and the relation of felonies to family income.

Graphs are invaluable because of the way they facilitate interpretation and analysis of both data and ideas. They enable the eye to take in at a glance important relationships that would be far less apparent from prose descriptions or long lists of numbers. But badly constructed graphs can confuse and mislead.

In this appendix we show how to read a graph that depicts a relationship between two variables, and we define the term *slope* and describe how it is measured and interpreted; finally, we explain how the behavior of three variables can be shown on a two-dimensional graph.

TWO-VARIABLE DIAGRAMS

Graphs are used to depict the behavior of economic variables such as number of jobs in the United States or the number of computers produced per year.

[1]Students who have a nodding acquaintance with geometry and feel quite comfortable with graphs can safely skip this appendix.

A **variable** is something measured by a number; it is used to analyze what happens to other things when the size of that number changes (varies).

Much of the economic analysis in this and other books requires that we keep track of two variables simultaneously. For example, in studying the operation of markets, we will want to keep one eye on the *price* of a commodity and the other on the *quantity* bought and sold.

For this reason, economists frequently find it useful to display real or imaginary figures in a *two-dimensional graph*, which simultaneously represents the behavior of two economic variables. The numerical value of one variable is measured along the bottom of the graph (called the *horizontal axis*), starting from the point labeled "0," and the numerical value of the other is measured up the side of the graph (called the *vertical axis*), also starting from "0."

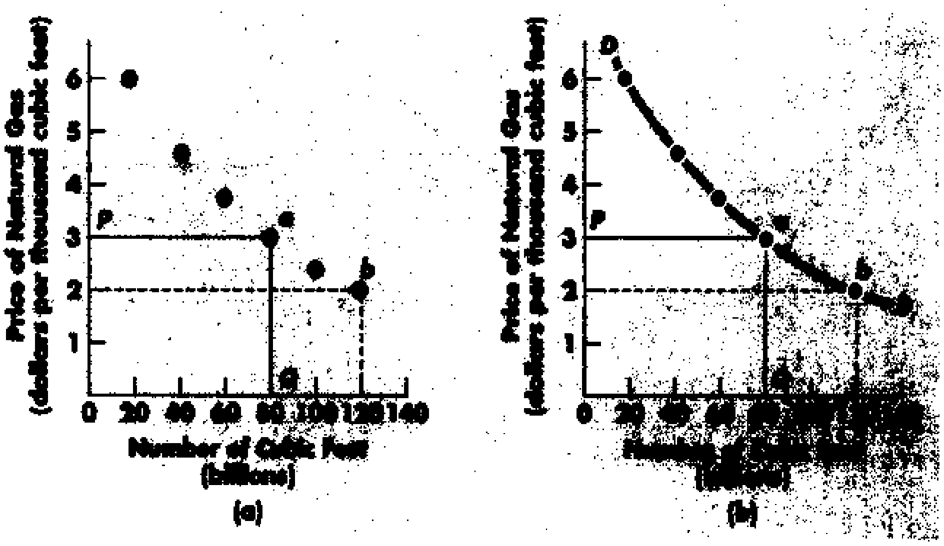
The "0" point in the lower left-hand corner of a graph where the axes meet is called the **origin.** Both variables are equal to zero at the origin.

Figures 1-5(a) and 1-5(b) are typical graphs of economic analysis. They depict an (imaginary) *demand curve*, represented by the orange dots in Figure 1-5(a) and the heavy orange line in Figure 1-5(b). The graphs show the price of natural gas on their vertical axes and the quantity of gas people want to buy at each such price on the horizontal axes. The dots in Figure 1-5(b) are connected by the continuous orange curve labeled *DD*.

Economic diagrams are generally read like latitudes and longitudes on a map. On the demand curve in Figure 1-5, point *a* represents a hypotheti-

FIGURE 1-5 A DEMAND CURVE FOR NATURAL GAS IN ST. LOUIS

This demand curve shows the relationship between the price of natural gas and the quantity of it that will be demanded. For example, the point labeled a indicates that, at a price of $3 per thousand cubic feet (point P), the quantity demanded will be 80 billion cubic feet (point Q).



cal combination of price and quantity demanded in St. Louis. By drawing a horizontal line leftward from that point to the vertical axis, we learn that the average price for gas in St. Louis is $3 per thousand cubic feet. By dropping a line straight down to the horizontal axis, we find that consumers want 80 billion cubic feet at this price, just as the statistics in Table 1-1 show. The other points on the graph give similar information. For example, point *b* indicates that if natural gas in St. Louis cost only $2 per thousand cubic feet, quantity demanded would be higher—it would reach 120 billion cubic feet.

Notice that information about price and quantity is *all* we can learn from the diagram. The demand curve will not tell us about the kinds of people who live in St. Louis, the sizes of their homes, or the conditions of their furnaces. It tells us about the price and the quantity demanded at that price; no more, no less. Specifically, it does tell us that when price declines there is an increase in the amount of gas consumers are willing and able to buy.

A diagram abstracts from many details, some of which may be quite interesting, in order to focus on the two variables of primary interest—in this case, the price of natural gas and the amount of gas that is demanded at each price. All the diagrams used in this book share this basic feature. They cannot tell the reader the "whole story" any more than a map's latitude and longitude figures for a particular city can make someone an authority on life in that city.

THE DEFINITION AND MEASUREMENT OF SLOPE

One of the most important features of the diagrams of economists is the steepness with which the line, or curve, being sketched runs uphill or downhill as we move to the right. The demand curve in Figure 1-5 clearly slopes downhill (the price falls) as we follow it to the right (that is, if more gas is to be demanded). In such instances we say that *the curve has a negative slope, or is negatively sloped, because one variable falls as the other one rises.*

The *slope of a straight line* is the ratio of the vertical change to the corresponding horizontal change as we move to the right along the line, or as it is often said, the ratio of the "rise" over the "run."

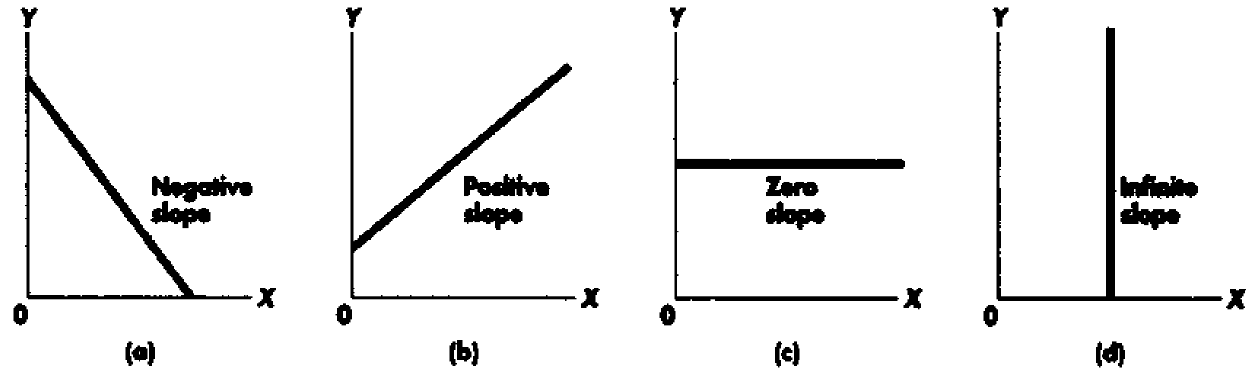
The four panels of Figure 1-6 show all the possible slopes for a straight-line relationship between two unnamed variables called Y (measured along the vertical axis) and X (measured along the horizontal axis). Figure 1-6(a) shows a negative slope,

TABLE 1-1

QUANTITIES OF NATURAL GAS DEMANDED AT VARIOUS PRICES

Price (per thousand cubic feet)	$2	$3	$4	$5	$6
Quantity demanded (billions of cubic feet)	120	80	56	38	20

FIGURE 1-6 DIFFERENT TYPES OF SLOPE OF A STRAIGHT-LINE GRAPH



In Figure 1-6(a), the curve goes downward as we read from left to right, so we say it has a negative slope. The slopes in the other figures can be interpreted similarly.

much like our demand curve. Figure 1-6(b) shows a positive slope, because variable *Y* rises (we go uphill) as variable *X* rises (as we move to the right). Figure 1-6(c) shows a *zero* slope, where the value of *Y* is the same irrespective of the value of *X*. Figure 1-6(d) shows an *infinite* slope, meaning that *X* does not have to change for *Y* to "leap through the roof."

Slope is a numerical concept, not just a qualitative one. The two panels of Figure 1-7 show two positively sloped straight lines with different slopes. The line in Figure 1-7(b) is clearly steeper. But by how much? The labels should help you compute the answer. In Figure 1-7(a) a horizontal movement, *AB*, of 10 units (13 − 3) corresponds to a vertical movement, *BC*, of 1 unit (9 − 8). So the slope is $BC/AB = 1/10$.

In Figure 1-7(b), the same horizontal movement of 10 units corresponds to a vertical movement of 3 units (11 − 8). So the slope is 3/10, which is larger.

By definition, the slope of any particular straight line is the same no matter where on that line we choose to measure it. That is why we can pick any horizontal distance, *AB*, and the corresponding slope triangle, *ABC*, to measure slope. But this is not true of lines that are curved.

Curved lines also have slopes, but the numerical value of the slope is different at every point.

The four panels of Figure 1-8 provide some examples of slopes of curved lines. The curve in Figure 1-8(a) has a negative slope everywhere, while the curve in Figure 1-8(b) has a positive slope everywhere. But these are not the only possibilities.

FIGURE 1-7 HOW TO MEASURE SLOPE

Slope indicates how much the graph rises per unit of movement from left to right. Thus, in Figure 1-7(b), as we go from point *A* to point *B*, we go 13 − 3 = 10 units to the right. But in that interval, the graph rises from the height of point *B* to the height of point *C*; that is, it rises 3 units. Consequently, the slope of the line is $BC/AB = 3/10$.

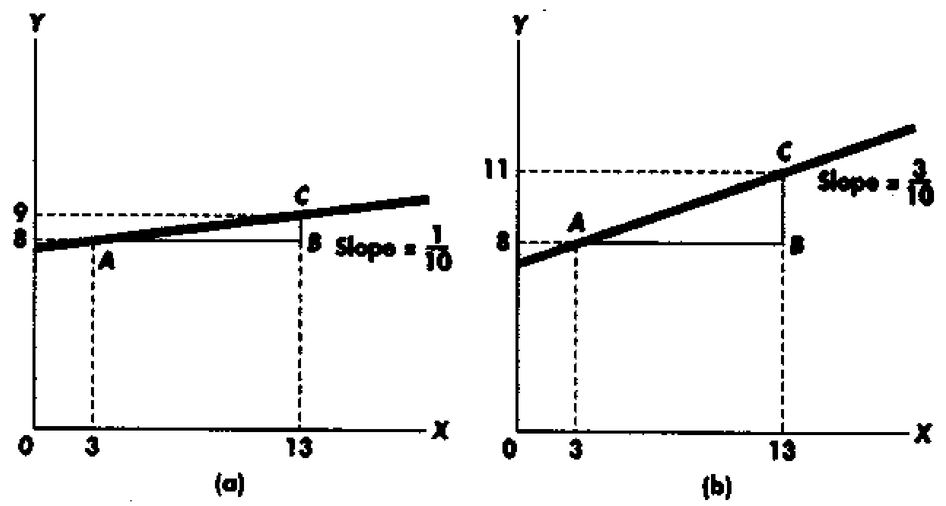
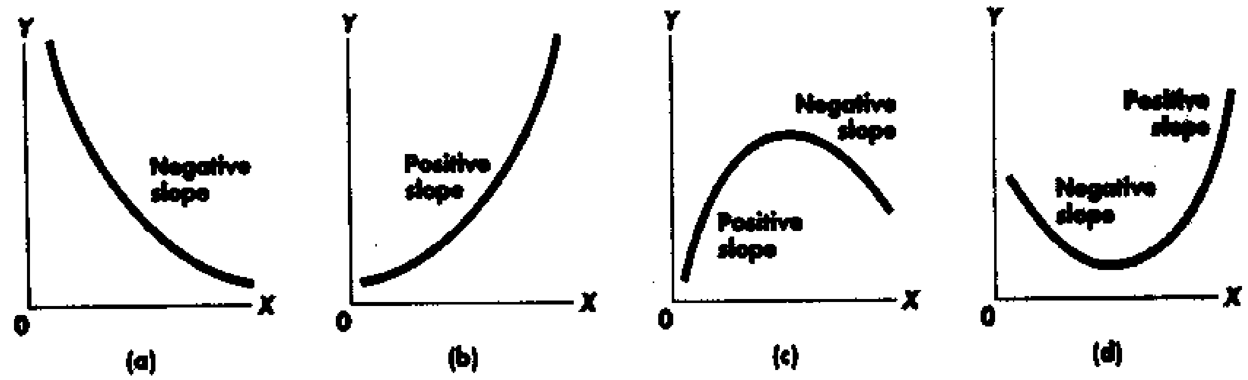


FIGURE 1-8 BEHAVIOR OF SLOPES IN CURVED GRAPHS



As Figures 1-8(c) and 1-8(d) indicate, where a graph is not a straight line it may have a slope that starts off as positive but that becomes negative farther to the right, or vice versa.

In Figure 1-8(c) we encounter a curve that has a positive slope at first but a negative slope later on. Figure 1-8(d) shows the opposite case: a negative slope followed by a positive slope.

It is possible to measure the slope of a smooth curved line numerically *at any particular point*. This is done by drawing a *straight* line that *touches*, but does not *cut*, the curve at the point in question. Such a line is called a *tangent* to the curve.

The slope of a curved line at a particular point is the slope of the straight line that is tangent to the curve at that point.

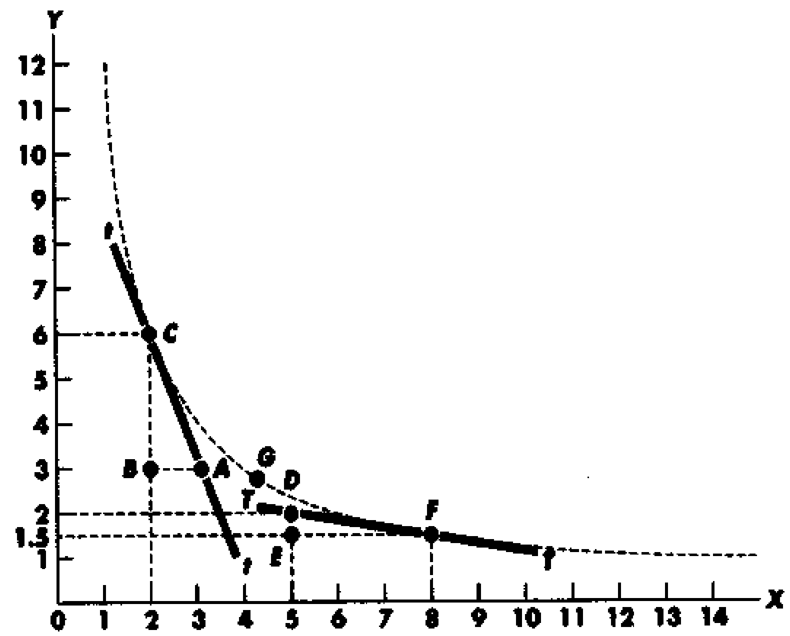
In Figure 1-9 we have constructed tangents to a curve at two points. Line *tt* is tangent at point *C*, and line *TT* is tangent at point *F*. We can measure the slope of the curve at these two points by applying the preceding definition. The calculation for point C, then, is the following:

$$\text{Slope at point C} = \text{Slope of line } tt = \frac{(\text{Distance } BC)}{(\text{Distance } BA)}$$

$$= \frac{(3-6)}{(3-2)} = \frac{(-3)}{(1)} = -3$$

FIGURE 1-9 HOW TO MEASURE SLOPE AT A POINT ON A CURVED GRAPH

To find the slope at point *F*, draw the line *TT*, which is tangent to the curve at point *F*; then measure the slope of the straight-line tangent *TT*, as in Figure 1-7. The slope of the tangent is the same as the slope of the curve at point *F*.



A similar calculation yields the slope of the curve at point F, which, as we can see from Figure 1-9, must be numerically smaller:

$$\text{Slope at point } F = \text{Slope of line } TT$$
$$= \frac{(1.5 - 2)}{(8 - 5)} = \frac{(-0.5)}{(3)} = -0.16$$

EXERCISE Show that the slope of the curve at point G is between −0.16 and −3.

What would happen if we tried to apply this graphical technique to the high point in Figure 1-8(c) or to the low point in Figure 1-8(d)? Take a ruler and try it. The tangents that you construct should be horizontal, meaning that they should have a slope exactly equal to zero. It is always true that where the slope of a smooth curve changes from positive to negative, or vice versa, there will be at least a single point with a zero slope.

Curves that have the shape of a hill, such as Figure 1-8(c), have a zero slope at their *highest* point. Curves that have the shape of a valley, such as Figure 1-8(d), have a zero slope at their *lowest* point.

RAYS THROUGH THE ORIGIN AND 45° LINES

The point at which a straight line cuts the vertical (Y) axis is called the *Y-intercept*. For example, the Y-intercept of the line in Figure 1-7(a) is a bit less than 8.

Lines whose Y-intercepts are zero have so many uses that they have been given a special name, a **ray through the origin**, or a ray.

Figure 1-10 contains three rays through the origin, and the slope of each is indicated in the diagram. The ray in the center—whose slope is 1—is particularly useful in many economic applications because it marks off points where X and Y are equal (as long as X and Y are measured in the same units). For example, at point A we have $X = 3$ and $Y = 3$, at point B, $X = 4$ and $Y = 4$, and a similar relation holds at any other point on that ray.

How do we know that this is always true for a ray whose slope is 1? If we start from the origin (where both X and Y are equal to zero) and the slope of the ray is 1, we know from the definition of slope that:

$$\text{Slope} = \frac{(\text{Vertical change})}{(\text{Horizontal change})} = 1$$

This implies that the vertical change and the horizontal change are always equal, so the two variables must always remain equal.

Rays through the origin with a slope of 1 are called **45° lines** because they form an angle of 45° with the horizontal axis. A 45° line marks off points

FIGURE 1-10 RAYS THROUGH THE ORIGIN

Rays are straight lines drawn through the zero point on the graph (the *origin*). Three rays with different slopes are shown. The middle ray, the one with slope = +1, has two very useful properties: (1) It makes a 45° angle with either axis, and (2) any point on that ray (for example, point A) is exactly equal in distance from the horizontal and vertical axes (length DA = length CA). Then at any point on that ray, such as A, the number on the X-axis (the abscissa) will be the same as the number on the Y-axis (the ordinate).

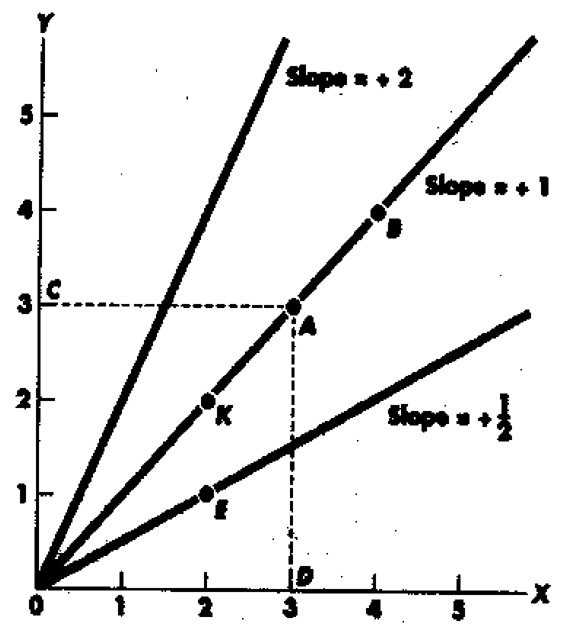
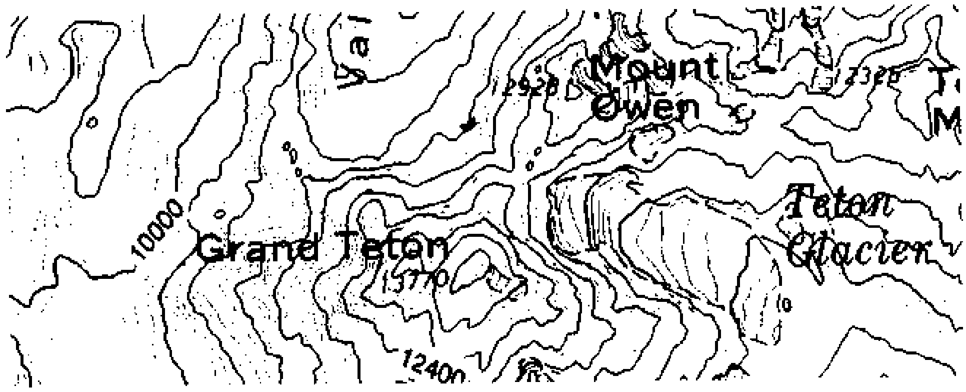


FIGURE 1-11 A GEOGRAPHIC CONTOUR MAP

All points on any particular contour line represent geographic locations that are at the same height above sea level.



where the variables measured on each axis have equal values.[2]

If a point representing some data is above the 45° line, we know that the value of Y exceeds the value of X. Correspondingly, whenever we find a point below the 45° line, we know that X is larger than Y.

SQUEEZING THREE DIMENSIONS INTO TWO: CONTOUR MAPS

Sometimes, because a problem involves more than two variables, two dimensions just are not enough, which is unfortunate since paper is only two dimensional. When we study the decision-making process of a business firm, for example, we may want to keep track simultaneously of three variables: how much labor the firm employs, how much raw material it imports from foreign countries, and how much output it creates.

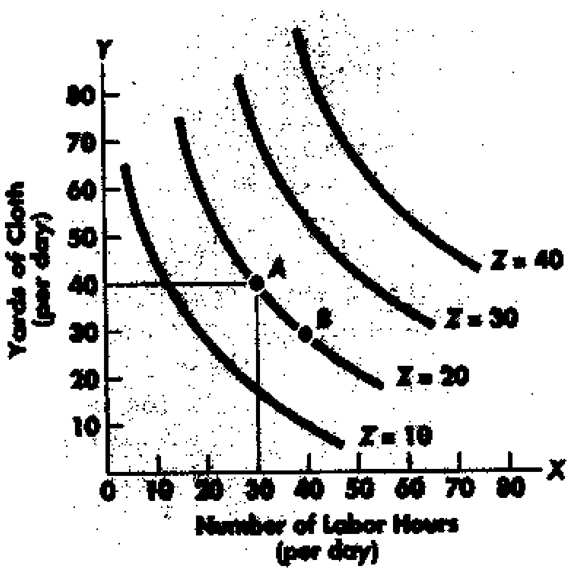
Luckily, there is a well-known device for collapsing three dimensions into two, namely a *contour map*. Figure 1-11 is a contour map of the Grand Teton mountains in the state of Wyoming. On the irregularly shaped "rings" we find numbers (such as 10,000 and 12,400) indicating the height above sea level at that particular spot on the mountain. Thus, unlike the more usual sort of map, which gives only latitudes and longitudes, this contour map exhibits three pieces of information about each point: latitude, longitude, and altitude.

Figure 1-12 looks more like the contour maps encountered in economics. It shows how some third variable, called Z (think of it as a firm's

[2]The definition assumes both variables are measured in the same units.

FIGURE 1-12 AN ECONOMIC CONTOUR MAP

In this contour map, all points on a given contour line represent different combinations of labor and raw materials capable of producing a given quantity of output. For example, all points on curve Z = 20 represent input combinations that can produce 20 units of output. Point *A* on that line means that the 20 units of output can be produced using 30 labor hours and 40 yards of cloth. Economists call such maps *production indifference maps*.



output, for example), varies as we change either variable X (think of it as a firm's employment) or variable Y (think of it as the use of imported raw material). Just like the map of the Grand Tetons, any point on the diagram conveys three pieces of data. At point A, we can read off the values of X and Y in the conventional way (X is 30 and Y is 40), and we can also note the value of Z by checking to see on which contour line point A falls. (It is on the $Z = 20$ contour.) So point A is able to tell us that 30 hours of labor and 40 yards of cloth produce 20 units of output.

While most of the analyses presented in this book will be based on the simpler, two-variable diagrams, contour maps will find their applications, especially in the appendixes to Chapters 5 and 7.

SUMMARY

1. Because graphs are used so often to portray economic models, it is important for students to acquire some understanding of their construction and use. Fortunately, the graphics used in economics are usually not very complex.
2. Most economic models are depicted in *two-variable diagrams*. We read data from these diagrams just as we read the latitude and longitude on a map: Each point represents the values of two **variables** at the same time.
3. In some instances, three variables must be shown at once. In these cases, economists use contour maps, which, as the name suggests, show "latitude," "longitude," and "altitude" all at the same time.
4. Often, the most important property of a line or curve drawn on a diagram will be its slope, which is defined as the ratio of the "rise" over the "run," or the vertical change divided by the horizontal change. Curves that go uphill as we move to the right have *positive slopes*, while curves that go downhill have *negative slopes*.
5. By definition, a *straight line* has the same *slope* wherever we choose to measure it. The *slope of a curved line* changes, but the slope at any point on the curve can be calculated by measuring the slope of a straight line *tangent to the curve* at that point.

KEY TERMS

Variable
Two-variable diagram
Horizontal and vertical axes
Origin (of a graph)
Slope of a straight (or curved) line
Negative, positive, zero, and infinite slope
Tangent to a curve
Y-intercept
Ray through the origin, or ray
45° line
Contour map

QUESTIONS FOR REVIEW

1. Look for a graph in your local newspaper, on the financial page or elsewhere. What does the graph try to show? Is someone trying to convince you of something with this graph?
2. Portray the following hypothetical data on a two-variable diagram:

Enrollment Data: University of Nowhere

Academic Year	Total Enrollment	Enrollment in Economics Courses
1990–1991	3,000	400
1991–1992	3,100	425
1992–1993	3,200	450
1993–1994	3,300	475
1994–1995	3,400	500

Measure the slope of the resulting line, and explain what this number means.

3. From Figure 1-9, calculate the slope of the curve at point G.
4. Sam believes that the number of job offers he will get depends on the number of courses in which his grade is B+ or better. He concludes from observation that the following figures are typical:

Number of grades of B+ or better	0	1	2	3	4
Number of job offers	2	4	5	6	7

Put these numbers into a graph like Figure 1-5(a). Measure and interpret the slopes between adjacent dots.

5. In Figure 1-10, determine the values of X and Y at point K and at point E. What do you conclude?
6. In Figure 1-12, interpret the economic meaning of points A and B. What do the two points have in common? What is the difference in their economic interpretation?

CHAPTER 2

THE U.S. ECONOMY: MYTH AND REALITY

E pluribus unum (Out of many, one)
Motto on U.S. currency

This chapter introduces you to the U.S. economy. It might seem that no such introduction is necessary, for you have probably lived your entire life in the United States. Every time you work at a summer or part-time job, pay your college bills, or buy a hot dog, you not only participate in the American economy, but observe something about it.

But the casual impressions we acquire in our everyday lives, while sometimes correct, are often misleading. Experience shows that most Americans—not just students—are either unaware of or harbor grave misconceptions about some of the most basic economic facts. One popular myth holds that America is inundated with imported goods, mostly from Japan. Another is that business profits account for something like a third of the price we pay for a typical good or service. Also, "everyone knows" that federal government jobs have grown rapidly over the last few decades. In fact, none of these things is true.

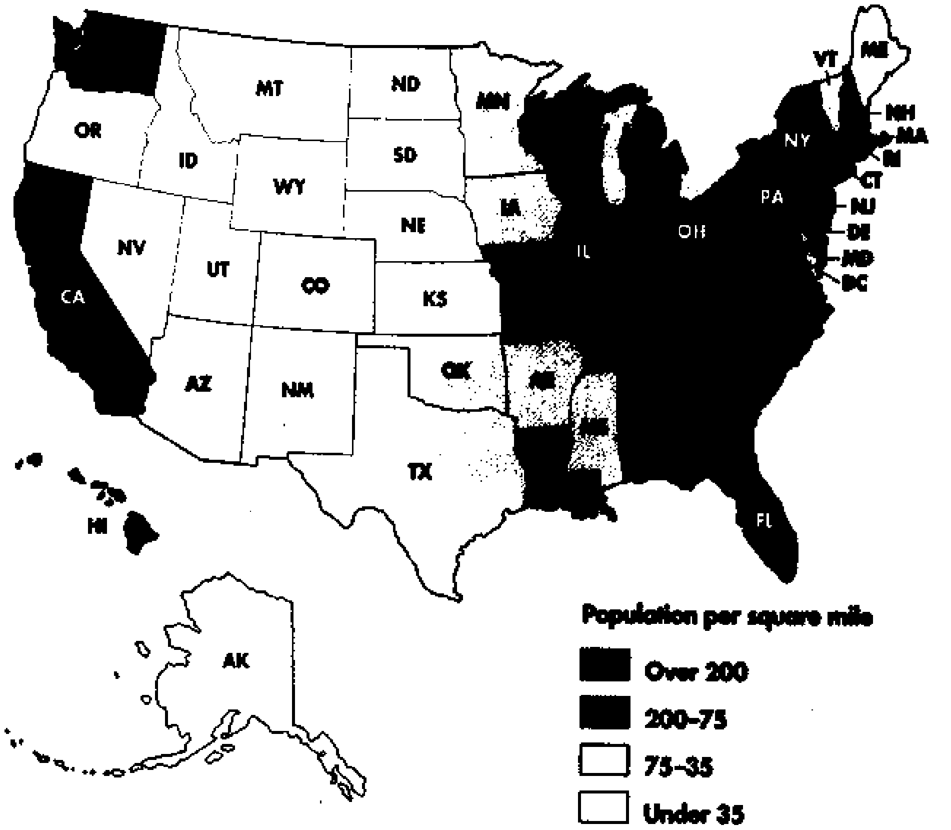
So, before we begin to study elaborate theories about how the economy works, it is useful to get a more *accurate* picture of what our economy looks like.

THE AMERICAN ECONOMY: A THUMBNAIL SKETCH

The U.S. economy is the biggest national economy on earth, for two very different reasons. First, there are a lot of us. The population of the United States is nearly 264 million—making it the third most populous nation on earth after China and India. That vast total includes children, retirees, full-time students, institutionalized people, and the unemployed, none of whom produce much output. But the *working population* of the United States numbers about 125 million. As long as they are reasonably productive, that many people are bound to produce vast amounts of goods and services. And they do.

But population is not the main reason why the U.S. economy is by far the world's biggest. After all, India has about three and a half times the population of the United States, but its economy is smaller than that of Texas. The second reason why the U.S. economy is so large is that we are a very rich country. Because American workers are among the most productive in the world, our

FIGURE 2-1 POPULATION DENSITIES BY STATE, 1994



This map divides the United States into four groups according to population density. The most densely populated states (over 200 people per square mile) are displayed in dark blue; they are clearly concentrated in the Northeast. States with intermediate population densities (75 to 200 people and 35 to 75 people per square mile, respectively) are shown in medium blue and light blue. They are found all over the nation, but especially in the Midwest and South. The most sparsely populated states, shown in white, are mainly in the West.

SOURCE: U.S. Bureau of the Census, *Statistical Abstract of the United States: 1995* (115th ed.), Washington, D.C.: U.S. Government Printing Office, 1995.

economy produces almost $28,000 worth of goods and services for every living American. If each of the 50 states was a separate country, California would be the sixth largest national economy on earth!

Why some countries (like the United States) are so rich and others (like India) are so poor is one of the central questions of economics. It is useful to think of an economic system as a machine which takes **inputs** such as labor and other things we call **factors of production** and transforms them into **outputs,** or the things people want to consume. The American economic machine performs this task with extraordinary efficiency, while the Indian machine runs quite inefficiently. Learning why is one of the chief reasons to study economics.

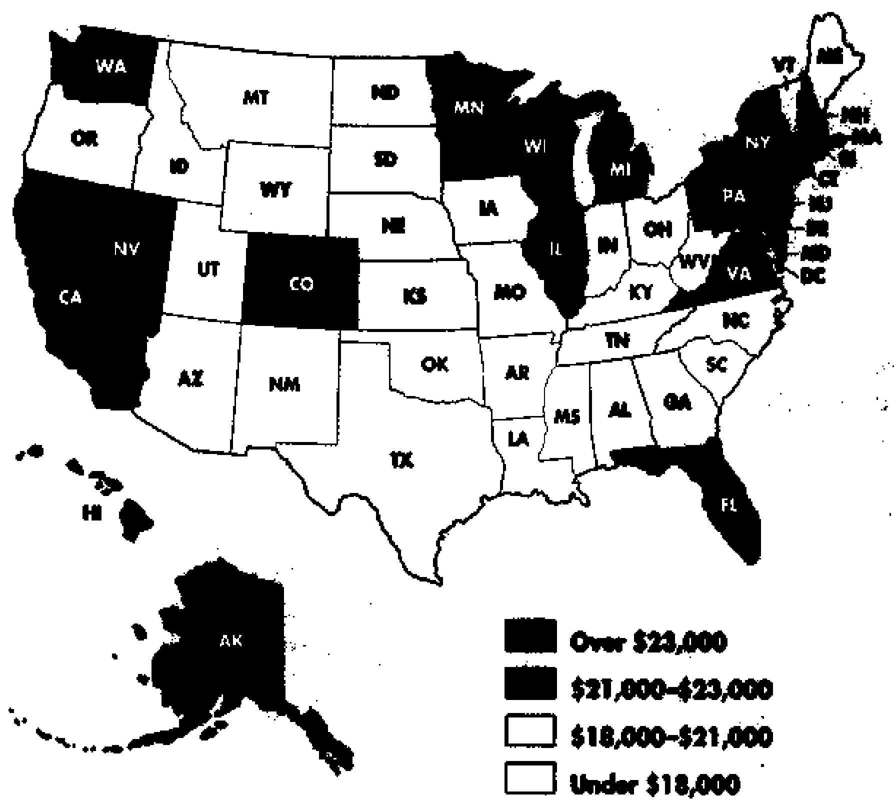
OUTPUTS are the goods and services that consumers want to acquire.

INPUTS or **FACTORS OF PRODUCTION** are the labor, machinery, buildings, and natural resources used to make outputs.

Thus, what makes the American economy the center of world attention is our unique combination of prosperity and population. There are other rich countries in the world, like Germany and Switzerland, and there are other countries with huge populations, like China and India. But no nation combines a huge population with high per-capita income the way the United States does. Japan, with an economy about 65 percent as large as ours, is the only nation that comes close.

While we are a rich and populous country, the 50 states certainly were not created equal. Population density varies enormously—from a high of more than 1,000 people per square mile in crowded New Jersey, to a low of just one person per square mile in Alaska. Figure 2-1 shows where population density is highest and lowest. Income variations are much less pronounced. But, still, the

FIGURE 2-2 DISPOSABLE INCOME PER CAPITA BY STATE, 1994



Although the United States is a unified economy, per-capita income nonetheless differs substantially from state to state. The richest states (incomes over $23,000 per capita), shown in dark blue, are concentrated in the Northeast, but also include California, Alaska, and Hawaii. The poorest states (per-capita incomes under $18,000, indicated in white, are found mainly in the South and West.

SOURCE: U.S. Bureau of the Census, *Statistical Abstract of the United States: 1995* (115th ed.), Washington, D.C.: U.S. Government Printing Office, 1995.

average income in Mississippi is only about half that in Connecticut. Figure 2-2 shows how per-capita income varies among the 50 states.

A PRIVATE-ENTERPRISE ECONOMY

Part of the secret of America's economic success is that free markets and private enterprise have flourished here. These days more than ever, private enterprise and capitalism are the rule, not the exception, around the globe. But the United States has taken the idea of the free market further than almost any other country. It remains "the land of opportunity."

Every country has a mixture of public and private ownership of property. Even in the darkest days of communism, Russians owned their own personal possessions. In our country, the post office and Tennessee Valley Authority are enterprises of the federal government, and many cities and states own and operate mass transit facilities and sports stadiums. But the United States stands out among the world's nations as among the most "privatized." Few industrial assets are publicly owned in America. Even many city bus companies, and almost all utilities (such as electricity, gas, and telephones), are run as private companies in the United States; in Europe, they are more often government enterprises.

GROSS DOMESTIC PRODUCT (GDP) is a measure of the size of an economy. It is, roughly speaking, the money value of all the goods and services produced in a year.

We are also one of the most "marketized" economies on earth. The standard measure of the total output of an economy is called **gross domestic product (or GDP)**, a term which appears frequently in the news. The share of GDP that passes through markets in the United States is enormous. While government

TABLE 2-1

OPENNESS OF VARIOUS NATIONAL ECONOMIES, 1994 (AVERAGE OF IMPORTS AND EXPORTS, AS A SHARE OF GDP)

The Netherlands	53%
Germany	30%
Canada	25%
United Kingdom	20%
Japan	13%
United States	**11%**
Mexico	10%
Russia	6%

SOURCE: Central Intelligence Agency, *The World Factbook 1995*, online version (http://www.odci.gov/cia/publications/95fact/index.html).

purchases of goods and services amount to about 17 percent of GDP, much of that is purchased from private businesses. Direct government *production* of goods is extremely rare in our society, and government services amount to only about 11 percent of GDP.

A RELATIVELY "CLOSED" ECONOMY

All nations trade with one another, and we are no exception. Our annual exports exceed $700 billion and our annual imports exceed $800 billion. That's a lot of money. But America's international trade often gets more attention than it deserves. The fact is that we produce most of what we consume and consume most of what we produce.

Among the most severe misconceptions about the U.S. economy is the myth that this country no longer manufactures anything, but rather imports everything from, say, Japan. In fact, only about 12 percent of America's GDP is imported, and only about a fifth of that comes from Japan. Contrary to a second myth, if we include *services* along with *goods* in the total, America's exports—at almost 11 percent of GDP—almost (but not quite) equal imports.

An economy is called relatively **OPEN** if its exports and imports constitute a large share of its GDP. An economy is considered relatively **CLOSED** if they constitute a small share.

Economists use the terms *open* and *closed* to indicate how important international trade is to a nation. A common measure of "openness" is the average of exports and imports, expressed as a share of GDP. Thus, the Netherlands is considered an extremely **open economy** because it imports and exports about 53 percent of its GDP. (See Table 2-1.) By this criterion, the United States stands out as among the most **closed economies** of the advanced, industrial nations. We export and import a smaller share of GDP than most of the countries listed in the table. Nonetheless, it is important to realize that this U.S. insularity is receding from the peak it attained just after World War II. U.S. exports rose from just 5.5 percent of GDP in 1972 to 9.0 percent in 1982 and 11 percent in 1994. We are increasingly becoming a great trading nation.

A GROWING ECONOMY . . . BUT WITH INFLATION

The next salient fact about the U.S. economy is its growth; it gets bigger almost every year. Gross domestic product in 1995 was $7.3 billion, more than 10 times as much as in 1965. Figure 2-3 charts the upward march of GDP in the United States since 1959. The rise looks impressive indeed; GDP increased more than 14-fold over the period. But the impression left by the graph is misleading. There is less here than meets the eye, and we need to understand why.

SOME PERILS IN READING GRAPHS: A DIGRESSION[1]

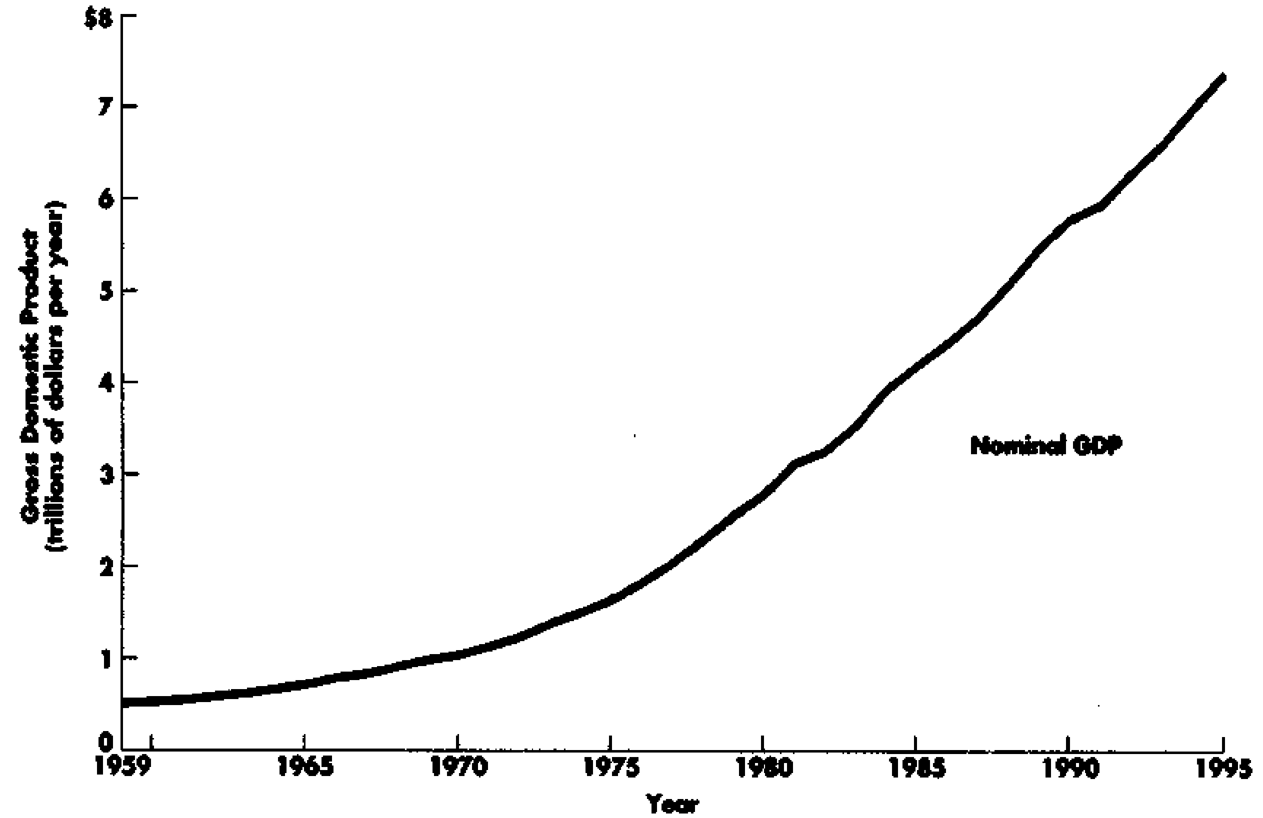
A **TIME-SERIES GRAPH** is a type of two-variable diagram in which time is the variable measured along the horizontal axis. It shows how some variable changed as time passed.

Graphs like Figure 2-3, in which time is measured horizontally and some economic variable (in this case GDP) is measured vertically, are often used to portray economic data. They are called **time-series graphs** and you will find a number of them in this book as well as in newspapers and magazines. Many of you will routinely read time-series charts in your work after college. So it is vital to understand what they do and do not show.

By summarizing an immense amount of data in compact form, time-series graphs can be invaluable—offering an instant visual grasp of the course of events.

[1]We will deal with some additional problems, beyond those discussed here, in an appendix to this chapter.

FIGURE 2-3 THE GDP OF THE UNITED STATES SINCE 1959



This time-series graph shows that the gross domestic product (GDP) of the United States, measured in dollars, has grown about 14-fold since 1959.

SOURCE: *Economic Report of the President 1996*, Washington, D.C.: U.S. Government Printing Office, 1996.

However, if misused, such graphs are very dangerous. They can easily mislead the inexperienced. Perhaps even more dangerous are the misconceptions perpetrated by people who draw graphs without sufficient care and, as a result, unintentionally mislead themselves and others. Consider Figure 2-3 as an example.

INFLATION refers to a sustained increase in the average level of prices.

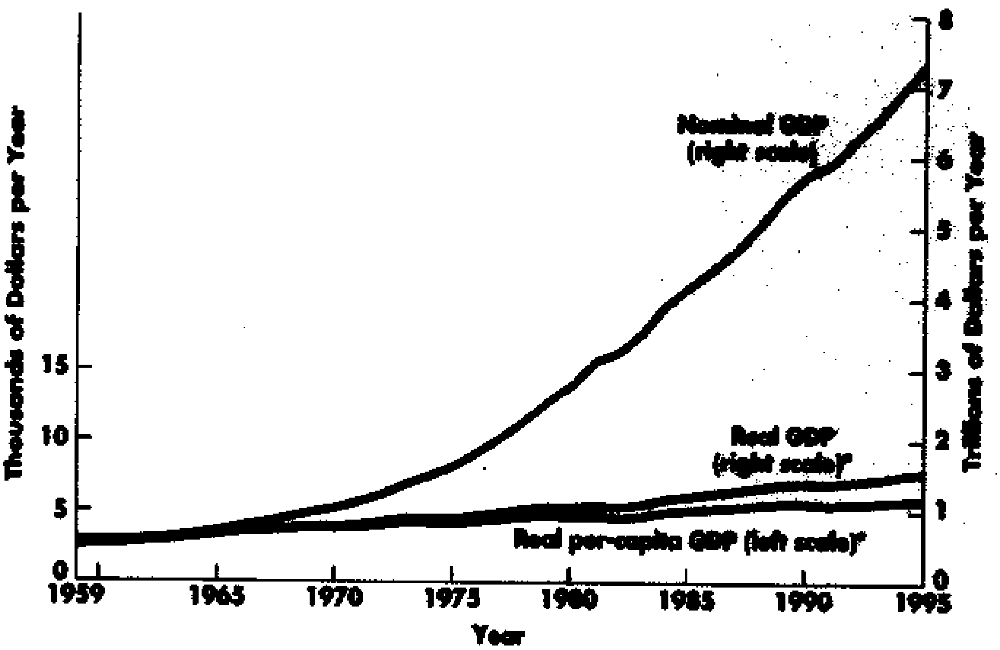
Most of the spectacular growth in GDP over this 36-year period was a reflection of two rather mundane facts. First, the price of almost everything rose between 1959 and 1995 because of **inflation**; in fact, average prices in 1995 were about five times higher than in 1959. Since each dollar in 1995 bought only about one-fifth of what it did in 1959, the dollar makes a rather poor measuring rod for comparing *production* in the two years. Most of the "growth" of GDP depicted in Figure 2-3 reflects inflation, not increases in output.

REAL GDP is the value of all the goods and services produced by an economy in a year, evaluated in dollars of constant purchasing power. Hence, inflation does not raise real GDP.

Economists correct for inflation by a process called *deflating by a price index*. Proper deflation leads to the orange line in Figure 2-4, which shows the growth of GDP *in dollars of constant purchasing power*. Economists call this **real GDP**. By this truer measure, we find that output in 1995 was about 3 times as high as in 1959, not 14 times as high.

Second, there were many more Americans alive in 1995 than in 1959—48 percent more to be exact. So output *per person* rose by considerably less than even the orange line in Figure 2-4 indicates. The blue line corrects for *both* inflation *and* population growth by charting the time-series behavior of GDP *per capita* in dollars of constant purchasing power. Americans were indeed richer in 1995 than

FIGURE 2-4 NOMINAL GDP, REAL GDP, AND REAL GDP PER CAPITA, 1959–1995



ª1959 dollars.

The orange line shows that real GDP, that is, GDP measured in dollars of constant purchasing power, has grown much less than GDP in current dollars (black line). Thus, most of the "growth" indicated by Figure 2-3 was just inflation. The blue line takes the next step and corrects for the fact that America was a much bigger place—in terms of population—in 1995 than in 1959. It therefore gives a more accurate picture of the increase in standards of living.

SOURCE: *Economic Report of the President 1996*, Washington, D.C.: U.S. Government Printing Office, 1996.

in 1959, but not by nearly as much as a naïve look at Figure 2-3 might suggest. In fact, the American standard of living roughly doubled over this period, rather than rising 14-fold. How misleading it can be simply to "look at the facts!" There is a general lesson to be learned from this example:

The facts, as portrayed in a time-series graph, most assuredly do not "speak for themselves." Because almost everything grows in a growing economy, one must use judgment in interpreting growth trends. Depending on what kind of data are being analyzed, and for what purpose, it may be essential to correct for population growth, for rising prices, or for other distorting or misleading influences.

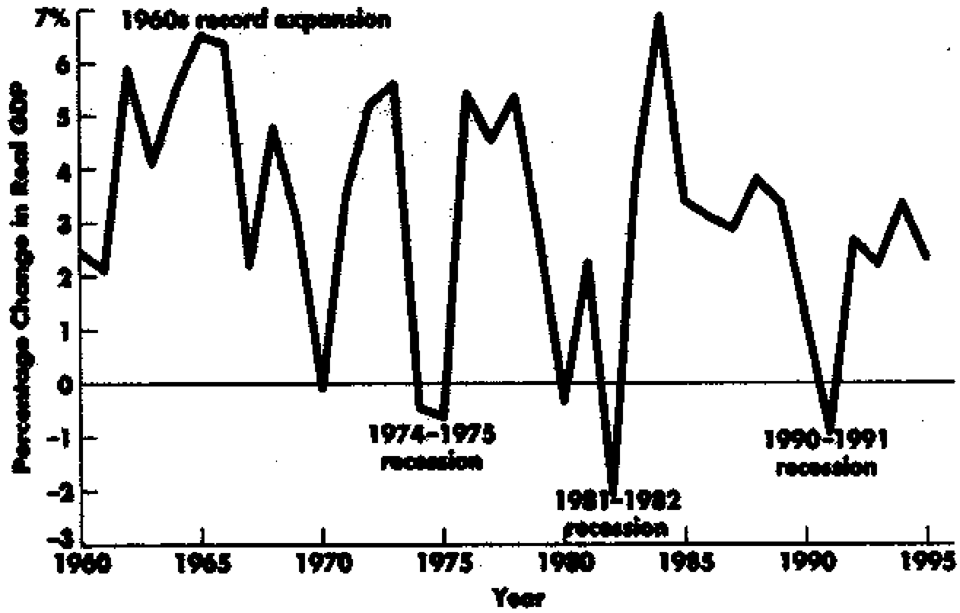
BUMPS ALONG THE GROWTH PATH: RECESSIONS

The bird's-eye view offered by Figure 2-4 conceals one more important fact: America's economic growth has been quite irregular. We have experienced alternating periods of good and bad times which are called *economic fluctuations* or sometimes just *business cycles*. In some years—five since 1959, to be exact—GDP actually declined. Such periods of *declining* economic activity are called **recessions**.

A **RECESSION** is a period of time during which the total output of the economy falls.

The bumps along the American economy's historic growth path are visible upon closer inspection of Figure 2-4, but stand out more clearly in Figure 2-5, which displays the same data in a different way. Here we plot not the *level* of real GDP each year, but rather its *growth rate*—the percentage change from one

FIGURE 2-5 THE GROWTH RATE OF REAL GDP IN THE UNITED STATES, 1960–1995



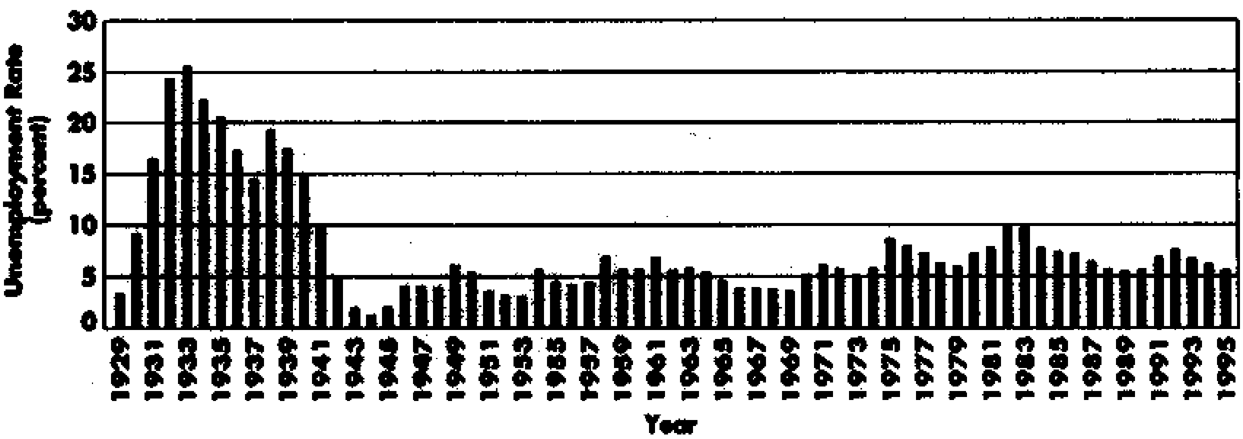
This diagram takes the same data used to construct the orange line in Figure 2-4, but uses it in a different way. Here we show the year-to-year growth rate (percentage increase) in real GDP from 1960 to 1995. Recessions stand out more clearly here as periods of *negative* growth.

SOURCE: *Economic Report of the President 1996*, Washington, D.C.: U.S. Government Printing Office, 1996.

year to the next. Now the booms and busts that delight and distress people—and swing elections—stand out clearly. From 1983 to 1984, for example, real GDP grew by close to 7 percent, which helped ensure the landslide reelection of Ronald Reagan. But from 1990 to 1991, real GDP actually fell by almost 1 percent, which helped Bill Clinton defeat George Bush.

One important consequence of these ups and downs in economic growth is that *unemployment* varies considerably from one year to the next. (See Figure 2-6.)

FIGURE 2-6 THE UNEMPLOYMENT RATE IN THE UNITED STATES, 1929–1995



This time-series chart shows the ups and downs of the U.S. civilian unemployment rate annually since 1929.

SOURCE: *Economic Report of the President 1996*, Washington, D.C.: U.S. Government Printing Office, 1996; and Bureau of the Census, *Historical Statistics of the United States, Colonial Times to 1970*, Washington, D.C.: U.S. Government Printing Office, 1975.

During the Great Depression of the 1930s, unemployment ran as high as 25 percent of the work force. But it fell to barely over 1 percent during World War II. Within the last few years, the national unemployment rate has been as low as 5.4 percent (in February 1995) and as high as 7.7 percent (in June 1992). In human terms, that 2.3 percentage point difference meant about 2.6 million *more* jobless workers. Understanding why joblessness varies so dramatically, and what we can do about it, is another major reason for studying economics.

THE INPUTS: LABOR AND CAPITAL

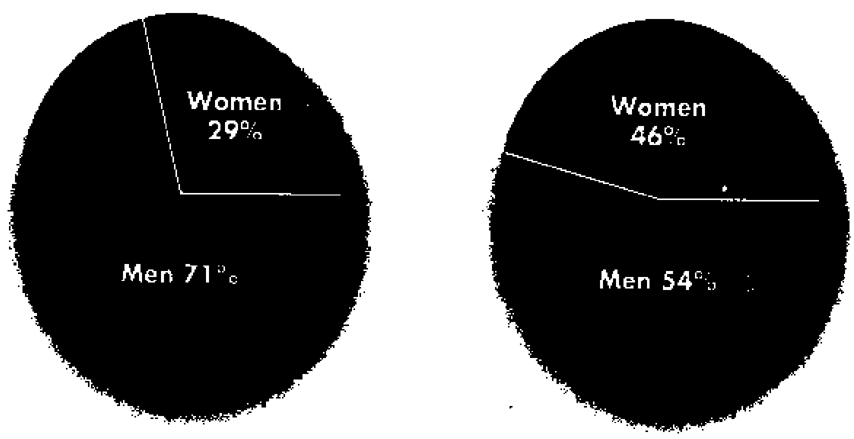
Let us now return to the analogy of an economy as a machine turning inputs into outputs. The most important input is human labor: the men and women who run the machines, work behind the desks, and serve you in the stores.

THE AMERICAN WORK FORCE: WHO IS IT?

We have already mentioned that about 125 million Americans hold jobs. Roughly 54 percent of these workers are men and 46 percent are women. This ratio represents a drastic change from a generation or two ago, when most women worked only at home. (See Figure 2-7.) Indeed, the massive entrance of women into the paid labor force has been one of the major social transformations of American life during the second half of the 20th century. In 1950, just 29 percent of women worked in the marketplace; now more than 46 percent do. The expanding role of women in the labor market has raised many controversial questions—such as whether they are discriminated against (the evidence suggests that they are), whether employers should be compelled to provide maternity leave, and so on.

In contrast to women, teenagers represent a dwindling share of the American work force. (See Figure 2-8.) Young men and women aged 16 to 19 accounted for 8.6 percent of employment in 1974 but only 5.0 percent in 1995. As the baby boom gave way to the baby bust, people under 20 became scarce resources! Still, over 6 million teenagers hold jobs in the U.S. economy today. Most are low-wage jobs at fast-food restaurants, amusement parks, and the like. Relatively few teenagers can be found in the nation's factories.

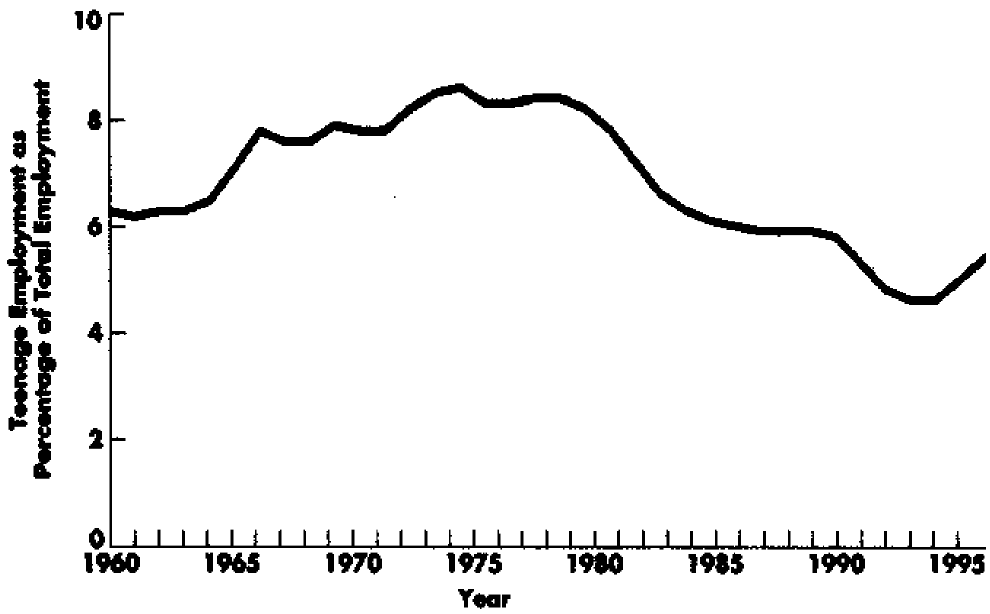
FIGURE 2-7 THE COMPOSITION OF EMPLOYMENT BY SEX, 1950 AND 1995



In 1950, just 29 percent of jobs were held by women. By 1995, this share had risen to 46 percent.

SOURCE: *Economic Report of the President 1996*, Washington, D.C.: U.S. Government Printing Office, 1996.

FIGURE 2-8 TEENAGE EMPLOYMENT AS A SHARE OF TOTAL EMPLOYMENT, 1960–1995



The share of teenagers (ages 16 to 19) in total employment rose from 6.3 percent in 1960 to a peak of 8.4 percent in 1974. Since then, it has largely been falling.

SOURCE: *Economic Report of the President 1996*, Washington, D.C.: U.S. Government Printing Office, 1996.

THE AMERICAN WORK FORCE: WHAT IT DOES

What do these 125 million people in the American work force do? The only real answer is: Almost anything you can imagine. In 1994, America had 656,000 lawyers, 559,000 bank tellers, 867,000 private security guards, 38,000 professional athletes, and 28,000 oil field roustabouts.

Figure 2-9 shows the breakdown by sector. It holds some surprises for most people. The majority of American workers—like workers in all developed countries—produce services, not goods. In 1994, about 90 million people were employed by service industries, including 27 million in retail and wholesale trade, while only 24 million produced goods. The popular image of the typical American worker as a factory hand—Homer Simpson, if you will—is really quite misleading. Manufacturing companies employ only about 18 million people, and almost a third of them work in offices rather than in the factory. Governments at all levels employ about 19 million people. Contrary to another popular misconception, few of these civil servants work for the *federal* government. Federal *civilian* employment is just under 3 million—and has fallen every year since 1990. Finally, about 3.6 million Americans work on farms, and the armed forces employ about 1.7 million soldiers.

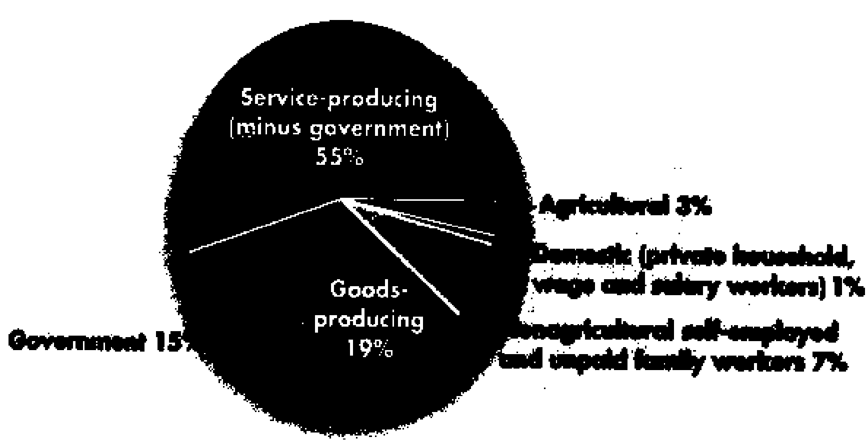
THE AMERICAN WORK FORCE: WHAT IT EARNS

All together, these workers earn nearly three-quarters of the income generated by the production process. That figures up to an average hourly wage of about $12—plus fringe benefits like health insurance and pensions, which can add an additional 30 to 40 percent for people holding what are often called "good jobs." Since the average work week is about 35 hours long, a typical weekly pay check

FIGURE 2-9 CIVILIAN EMPLOYMENT BY SECTOR, 1994

More Americans produce services than goods. In fact, all levels of government now employ more workers than the manufacturing sector.

SOURCE: "Current Labor Statistics," *Monthly Labor Review*, Vol. 118, No. 11, November 1995, p. 47.



is about $420 before taxes. That is hardly a princely sum, and most college graduates can expect to earn more.[2] But that is what wage rates are like in a rich country. Wages in Japan and throughout northern Europe are similar.

CAPITAL AND ITS EARNINGS

After deducting the tiny sliver of income that goes to land and natural resources, most of the remainder accrues to the owners of *capital*—the machines and buildings that make up the nation's industrial plant.

The total market value of American business assets—a tough number to estimate—is believed to be in the neighborhood of $12 trillion. Since that capital earns an average rate of return of about 10 percent before taxes, the total earnings of capital come to about $1,200 billion. Of this, profits are less than half; the rest is mainly interest.

Public opinion polls routinely show that Americans have a distorted view of the level of business profits in our society. The man and woman in the street believe that profits account for about 30 percent of the price of a typical product. In fact, when you spend a dollar in our economy, about 66 cents is for labor costs, 11 cents goes to cover the wear and tear on the capital stock,[3] 13 cents is for taxes, and 4 cents is for interest. That leaves about 6 cents for after-tax profits.

THE OUTPUTS: WHAT DOES AMERICA PRODUCE?

What does all this labor and capital produce? Consumer spending accounts for almost 70 percent of GDP. And what an amazing variety of goods and services it buys! American households spend roughly half of their budgets on goods—ranging from $93 billion per year on new cars to $51 billion on tobacco products. Expenditures on services absorb the other half of household budgets, of which housing commands the largest share. But Americans also spend $68 bil-

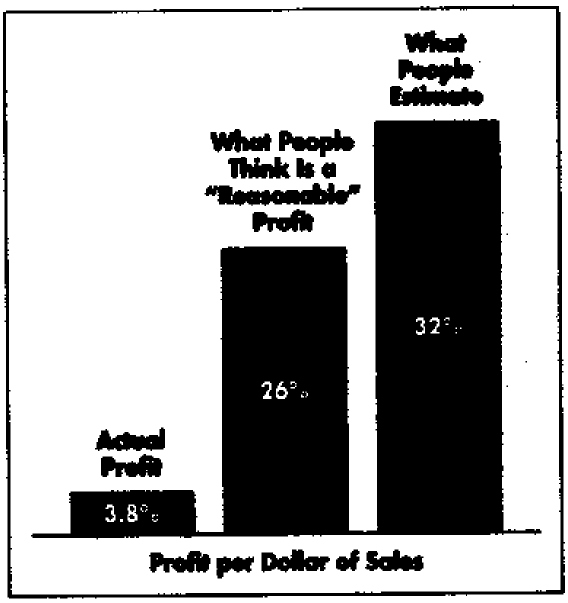
[2]These days, college graduates typically earn about 65 percent more than those with only high school diplomas.

[3]Economists and accountants call this *depreciation*. It is a well-known cost of doing business.

PUBLIC OPINION ON PROFITS

Most Americans think corporate profits are much higher than they actually are. One public opinion poll, for example, found that the average citizen thought that corporate profits *after tax* amounted to 32 percent of sales for the typical manufacturing company. The actual profit rate at the time was closer to 4 percent! Interestingly, when a previous poll asked how much profit was "reasonable," the response was 26 cents on every dollar of sales—over six times as large as profits actually were.

SOURCE: "Public Attitudes toward Corporate Profits," Opinion Research Corporation, *Public Opinion Index*, Princeton, N.J., June 1986.



lion annually on their telephone bills, $29 billion on airline tickets, and $39 billion on dentists.

This leaves close to a third of GDP for all nonconsumption uses. The government buys about 17 percent: such things as airplanes, guns, and the services of soldiers, teachers, and bureaucrats. The rest is mainly business purchases of machinery and industrial structures (about 10 percent of GDP) and consumer purchases of new houses (about 4 percent).

THE CENTRAL ROLE OF BUSINESS FIRMS

Calvin Coolidge once said that "the business of America is business." He was largely right. When we peer inside the economic machine that turns inputs into outputs, we see mainly private companies. Astonishingly, the United States has about 21 million business firms—about one for every 13 people!

The owners and managers of these businesses hire people, acquire or rent capital goods, and arrange for the production of the things people want to buy. Sound simple? It isn't. About 75,000 to 80,000 businesses fail every year. A few succeed spectacularly. Some do both. Wang Laboratories, an early entrant into the calculator and word processing business, was founded by a brilliant Chinese immigrant in the 1950s and grew to have annual sales of $3 billion and 30,000 employees by 1984. Eight years later, it was bankrupt. The company's financial condition has improved since 1993, but it is still losing money. Fortunately for the U.S. economy, however, the lure of riches induces hundreds of thousands of people to start new businesses every year—against the odds.

A number of the biggest firms do business all over the world, just as foreign-based *multinational corporations* do business here. Indeed, some people claim that it is now impossible to determine the true "nationality" of a multinational

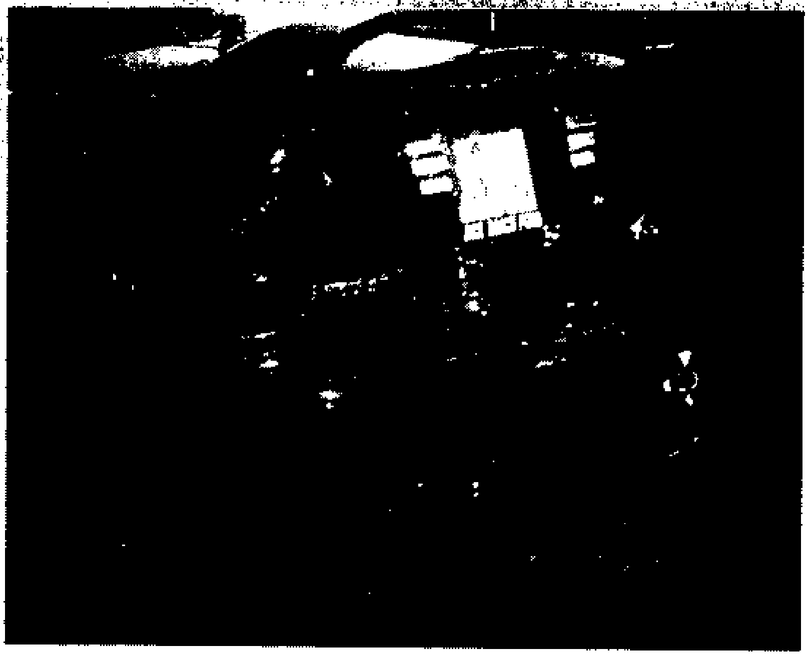
IS THAT AN AMERICAN COMPANY?

Robert Reich, the secretary of labor in the Clinton administration, has argued that it is almost impossible to define the nationality of a multinational company these days. While many scholars think Reich exaggerates the point, no one doubts that he has one.

Here are some examples:

> What's the difference between an "American" corporation that makes or buys abroad much of what it sells around the world and a "foreign" corporation that makes or buys in the United States much of what it sells? . . . The mind struggles to keep the players straight. In 1990, Canada's Northern Telecom was selling to its American customers telecommunications equipment made by Japan's NTT at NTT's factory in North Carolina.
>
> If you found that one too easy, try this: Beginning in 1991, Japan's Mazda would be producing Ford Probes at Mazda's plant in Flat Rock, Michigan. Some of these cars would be exported to Japan and sold there under Ford's trademark.
>
> A Mazda-designed compact utility vehicle would be built at a Ford plant in Louisville, Kentucky, and then sold at Mazda dealerships in the United States. Nissan, meanwhile, was designing a new light truck at its San Diego, California, design center. The trucks would be assembled at Ford's Ohio truck plant, using panel parts fabricated by Nissan at its Tennessee factory, and then marketed by both Ford and Nissan in the United States and in Japan. Who is Ford? Nissan? Mazda? . . .

SOURCE: Robert B. Reich, *The Work of Nations* (New York: Knopf, 1991), pp. 124, 131.



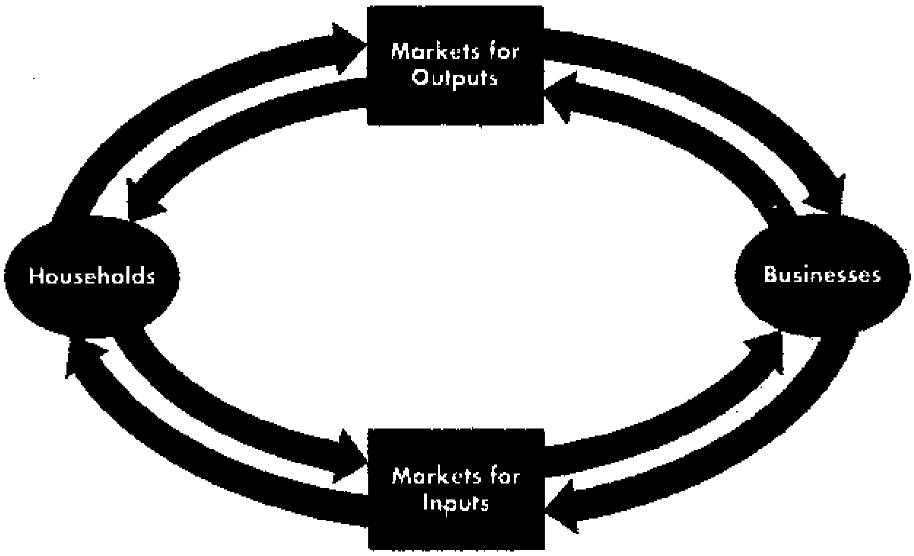
corporation—which may have factories in ten or more countries, sell its wares all over the world, and have stockholders in dozens of nations. (See the accompanying box.) Most of General Motors' profits are generated abroad, for example. And the Honda you drive was probably made in Ohio.

Firms compete with other companies in their *industry*. Many economists believe that this *competition* is the key to industrial efficiency. The sole supplier of a commodity will find it easy to make money, and may therefore fail to innovate or control costs. Its management is liable to become relaxed and sloppy. But a company besieged by dozens of competitors eager for its business must keep alert at all times. The rewards for success in business can be magnificent. But the punishment for failure is severe.

WHAT'S MISSING FROM THE PICTURE? GOVERNMENT

Thus far we have the following capsule summary of how the U.S. economy works: About 21 million private businesses, energized by the profit motive, employ about 125 million workers and about $12 trillion of capital. These firms bring their enormously diverse wares to market, where they try to sell them to nearly 264 million consumers.

FIGURE 2-10 THE CIRCULAR FLOW OF GOODS AND MONEY



This diagram depicts the ways in which households and firms interact in markets. The upper half signifies the markets for outputs—the goods and services that firms produce and households purchase. The lower half illustrates the markets for inputs like labor and capital that households sell (or rent) to businesses. In each case, the outer loop indicates the flow of money while the inner loop signifies the flow of physical goods and services.

Households and businesses are linked together in a tight circle, depicted in Figure 2-10. Firms use their receipts from sales to pay wages to their employees and interest and profits to the people who provide them with capital. These income flows, in turn, enable consumers to purchase the goods and services that companies produce. This circular flow of money and goods is central to the analysis of how the national economy works. All these activities are linked together by a series of interconnected markets, some of which are highly competitive and others of which are less so.

All very well and good. But the story leaves out something important: the role of *government*, which is pervasive even in our decidedly free-market economy. Just what does government do in the U.S. economy—and why?

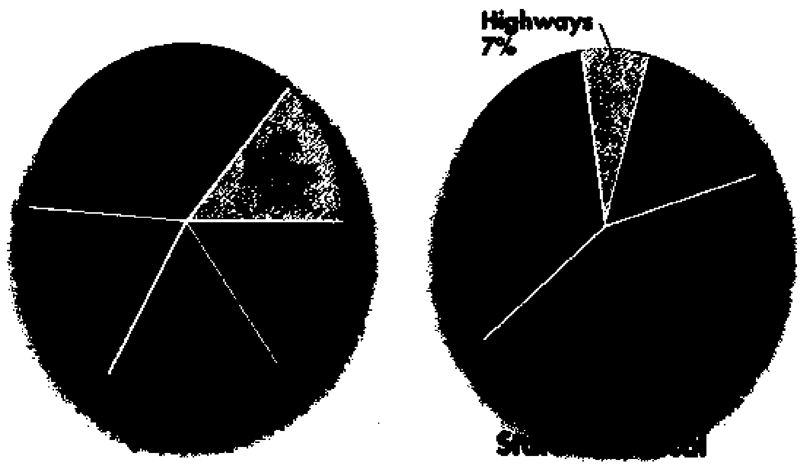
While an increasing number of tasks seem to get assigned to the state each year, the traditional role of government in a market economy revolves around five jobs: providing certain goods and services such as national defense, levying taxes to pay for these goods and services, redistributing income, regulating business, and making and enforcing the laws. Every one of these is steeped in controversy and surrounded by intense political debate. We conclude this chapter with a brief look at the role of government.

THE GOVERNMENT AS REFEREE

For the most part, power is diffused in our economy, and people "play by the rules." But, in the scramble for competitive advantage, disputes are bound to arise. Did Company A live up to its contract? Who owns that disputed piece of property? In addition, some unscrupulous businesses are liable to step over the line now and then—via misleading advertising, attempts to monopolize markets, employment of child labor, and the like.

Enter the government as rule maker, referee, and arbitrator. Congress and state and local legislatures pass the laws that define the rules of the economic game. The executive branches of all three levels of government have the responsibility for enforcing them. And the courts interpret the laws and adjudicate disputes.

FIGURE 2-11 THE ALLOCATION OF GOVERNMENT EXPENDITURES



These graphs show how the government spends tax dollars. The federal government spends most of its money on transfer payments to retirees, the poor, the unemployed, and veterans (37 percent) and on national defense (18 percent). The biggest share of state and local government spending goes for education (34 percent), with health and welfare expenditures (16 percent) in second place.

SOURCE: *Economic Report of the President 1996*, Washington, D.C.: U.S. Government Printing Office, 1996.

REGULATING BUSINESS

Nothing is pure in this world of ours. Even in "free-market" economies, governments interfere with the workings of free markets in many ways and for myriad reasons. Some regulations grow out of the rule-making function. For example, America's *antitrust laws* are designed to protect competition against possible encroachment by monopoly. Others are aimed at promoting social objectives to which unfettered markets do not tend; environmental regulations are a particularly clear case. But some economic regulations have *no* persuasive economic rationale at all!

We mentioned earlier that the American belief in free enterprise runs deep. For this reason, the regulatory role of government is more contentious here than in most other countries. It was, after all, Jefferson who said that government is best which governs least. Two hundred years later, Presidents Reagan, Bush, and Clinton pledged to dismantle regulations—and sometimes did.

GOVERNMENT EXPENDITURES

The most contentious political issues often involve taxing and spending because those are the government's most prominent roles. In 1995 and 1996, President Clinton and the Republican-led Congress battled fiercely over the federal budget. Differences over tax policies and spending cuts even led to some temporary shutdowns of the government.

During fiscal year 1995, the federal government spent about $1.5 *trillion*—a sum that is literally beyond comprehension. Figure 2-11 shows where the money went. Over one-third went for *pensions and income security* programs, which include both social insurance programs, like Social Security and unemployment compensation, and programs designed to assist the poor. About 18 percent went for *national defense*. Another 18 percent was absorbed by *health-care* expenditures, mainly on Medicare and Medicaid. Adding in *interest on the national debt*, these four functions alone accounted for about 88 percent of federal spending. The rest went for a miscellany of other purposes including education, transportation, agriculture, housing, and foreign aid.

Government spending at the state and local levels was about $1.4 trillion. Education claimed the biggest share of state and local government budgets (34 percent), with health and public welfare programs in second place (16 percent).

Despite this vast outpouring of public funds, many observers believe that serious social needs remain unmet. Critics claim that our public infrastructure (such as bridges and roads) is inadequate, that our educational system is lacking, that we do not do enough for the poor and homeless, and so on. Many of these claims were echoed during the 1995 budget debate. Other critics argue that government tries to do too much—and does it too inefficiently.

TAXES IN AMERICA

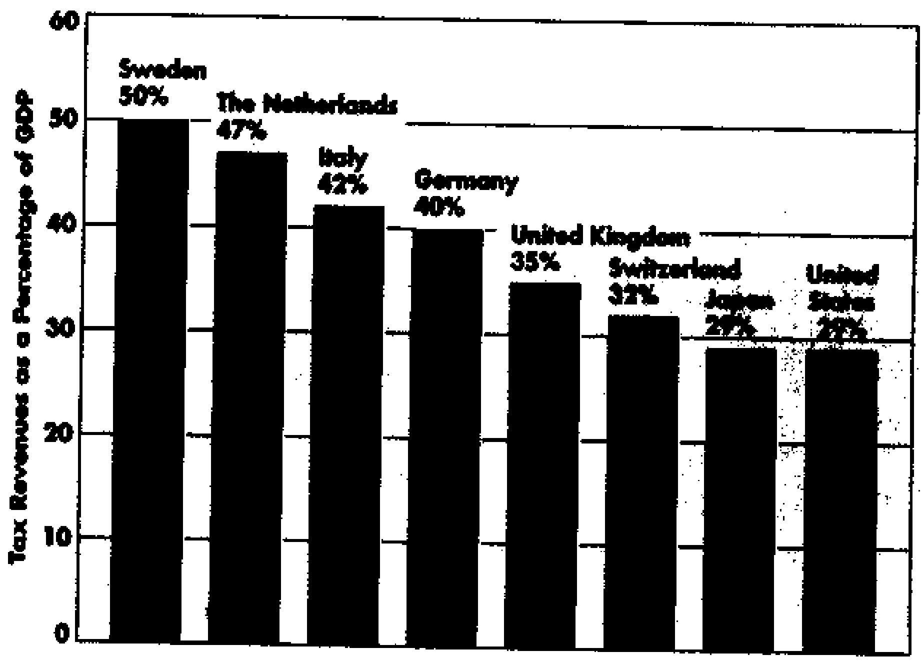
Taxes are required to finance this array of goods and services, and sometimes it seems that the tax collector is everywhere. We have income and payroll taxes withheld from our paychecks, sales taxes added to our purchases, property taxes levied on our homes; we pay gasoline taxes, liquor taxes, and telephone taxes.

Americans have always felt that there are too many taxes and that they are too high. In the 1980s and 1990s, antitax sentiment became a dominant feature of the U.S. political scene. The old slogan "no taxation without representation" gave way to the new slogan "no new taxes." Yet by international standards, Americans are among the most lightly taxed people in the world. Figure 2-12 compares the fraction of income paid in taxes in the United States with those paid by residents of other wealthy nations. Americans shoulder some of the lowest tax burdens in the industrialized world.

How is this money raised? The *personal income tax* is the federal government's biggest revenue source, though the *payroll tax*—a flat-rate tax on wages and salaries up to a certain limit—is not far behind. The rest of federal revenue comes mainly from the *corporate income tax*. Most states and many large cities levy

FIGURE 2-12

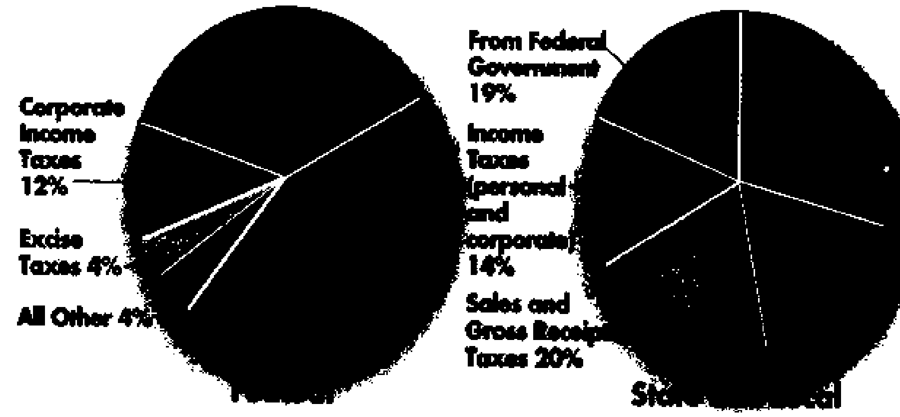
THE BURDEN OF TAXATION IN SELECTED COUNTRIES, 1992



Americans are lightly taxed compared to citizens of other advanced industrial countries. The Swedes and the Dutch, for example, pay far higher taxes than we do.

SOURCE: U.S. Bureau of the Census, *Statistical Abstract of the United States: 1995* (115th ed.), Washington, D.C.: U.S. Government Printing Office, 1995.

FIGURE 2-13 SOURCES OF GOVERNMENT REVENUE



These pie diagrams show the shares of each of the major sources of federal and state and local revenues. Personal income taxes and payroll taxes clearly account for the majority of federal revenues. The states and localities raise money from a potpourri of sources, including the federal government.

SOURCE: *Economic Report of the President 1996*, Washington, D.C.: U.S. Government Printing Office, 1996.

broad-based *sales taxes* on retail purchases, with certain specific exemptions (such as food and rent). Local governments generally raise revenue by levying *property taxes* on homes and business properties. Figure 2-13 shows the breakdown of revenue sources at both the federal and state and local levels.

THE GOVERNMENT AS REDISTRIBUTOR

In a market economy, people earn incomes according to what they have to sell. Unfortunately, many people have nothing to sell but unskilled labor, which commands a paltry price. Others lack even that. Such people fare poorly in unfettered markets. In extreme cases, they are homeless, hungry, and ill. Robin Hood transferred money from the rich to the poor. Some think the government should do the same; others disagree.

If poverty amidst riches offends your moral sensibilities—a personal judgment that each of us must make—there are two basic approaches. The socialist idea is to force the distribution of income to be more equal by overriding the workings of markets. "From each according to his ability, to each according to his needs" was Marx's ideal. In practice, things were not quite so noble under socialism. But there is little doubt that incomes in the old Soviet Union were more equally distributed than those in the United States.

TRANSFER PAYMENTS are sums of money that certain individuals receive as outright grants from the government rather than as payments for services rendered.

A tax is **PROGRESSIVE** if the ratio of taxes to income rises as income rises.

The liberal idea is to let free markets determine the distribution of *before-tax* incomes, but then to use the tax system and **transfer payments** to reduce inequality—just as Robin Hood did. This is the rationale for, among other things, **progressive taxation** and the antipoverty programs colloquially known as "welfare." Americans who support redistribution line up solidly behind the liberal approach. But which ways are the best, and how much is enough? There are no simple answers to these highly contentious questions.

CONCLUSION: THE MIXED ECONOMY

Ideology notwithstanding, all nations at all times blend public and private ownership of property in some proportions. All rely on markets for some purposes; but all also assign some role to government. Hence, people speak of the ubiquity of **mixed economies.** But mixing is not homogenization; different countries can and do blend the state and market sectors in different ways. Even today, the

A **MIXED ECONOMY** is one in which there is some public influence over the workings of free markets. There may also be some public ownership mixed in with private property.

Russian economy is a far cry from the Italian economy, which is vastly different from that of Hong Kong.

While most of you were in middle school, a stunning historical event occurred: Communism collapsed all over Europe. Now the formerly socialist economies are in the midst of a painful transition from a system in which private property, free enterprise, and markets played subsidiary roles to one in which they are central. These nations are changing the mix, if you will—and dramatically so. To understand why this transformation is at once so difficult and so important, we need to explore the main theme of this book: *What does the market do well, and what does it do poorly?* This task begins in the next chapter.

SUMMARY

1. The U.S. economy is the biggest national economy on earth, both because Americans are rich by world standards and because we are a populous nation. Relative to most other advanced countries, our economy is also exceptionally "privatized" and *closed*.
2. The U.S. economy has grown dramatically over the years. But this growth is exaggerated by looking only at dollar figures, which are distorted by both **inflation** and population growth. To get a better understanding of the growth of living standards, we must look at **real GDP** *per capita*.
3. The growth path of the U.S. economy has been interrupted by periodic **recessions**, during which unemployment rises.
4. America has a big, diverse work force whose composition by age and sex has been changing substantially. Relatively few workers these days work in factories or on farms; most work in service industries.
5. Employees take home most of the nation's income. Most of the rest goes, in the forms of interest and profits, to those who provide the capital.
6. Governments at the federal, state, and local levels employ almost one-seventh of the American work force and produce almost a fifth of the GDP. They finance their expenditures by taxes, which account for about 29 percent of GDP. This percentage is one of the lowest in the industrialized world.
7. In addition to raising taxes and making expenditures, the government in a market economy serves as referee and enforcer of the rules, regulates business in a variety of ways, and redistributes income through taxes and **transfer payments**. For all these reasons, we say that we have a **mixed economy** which blends private and public elements.

KEY TERMS

Outputs
Inputs (factors of production)
Gross domestic product (GDP)
Time-series graph
Inflation
Real GDP
Open economy
Closed economy
Recession
Transfer payments
Progressive tax
Mixed economy

QUESTIONS FOR REVIEW

1. Which are the two biggest national economies on earth? Why are they so much bigger than the others?
2. What is meant by a "factor of production?" Have you ever sold any on a market?
3. Do you have any ideas why per-capita income in Connecticut is nearly double that in Mississippi?
4. What is the difference between nominal gross domestic product and real gross domestic product? Why is this distinction important?
5. Roughly speaking, what fraction of U.S. labor works in factories? In service businesses? In government?
6. It sounds paradoxical to say that most American businesses are small, but most of the output is produced by large businesses. How can this be true?
7. What is the role of government in a mixed economy?

APPENDIX FURTHER PERILS IN THE INTERPRETATION OF GRAPHS

The chapter warned you against certain dangers that arise in interpreting time-series graphs. But there are other perils as well. This appendix deals with three of them.

DISTORTING TRENDS BY CHOICE OF THE TIME PERIOD

Users of statistical data must be on guard for distortions of trends caused by an unskillful or unscrupulous choice of the beginning and ending time periods for the graph. This is best explained by an example.

Figure 2-14 shows the behavior of average stock market prices over the period January 1966 to June 1982. The numbers have been corrected for inflation; that is, they are expressed in dollars of constant purchasing power. The graph displays a clear downhill movement that would suggest to anyone not familiar with other information that stocks are a terrible investment.

However, an unscrupulous seller of stocks could use similar stock market statistics for a different group of years to tell exactly the opposite story. Figure 2-15 shows the behavior of average stock prices from 1989 to 1990. Stocks now look like a superb investment.

A much longer and less biased choice of time period (Figure 2-16) gives a less distorted picture. It indicates that investments in stocks are sometimes profitable, sometimes unprofitable. The lesson is that:

You must be on the lookout for distortions resulting from an inadvertent or unscrupulous choice of time period for a graph.

While no ironclad rules can give absolute protection from this difficulty, several precautions can be helpful.

1. Make sure the first date on the graph is not an exceptionally high or low point. Compared to 1966, a year of unusually high stock prices, the years immediately following are bound to give the impression of a downward trend.
2. For the same reason, make sure the graph does not end in a year that is extraordinarily high or

FIGURE 2-14 STOCK PRICES, JANUARY 1966–JUNE 1982

This graph seems to show that stock market prices generally go down.

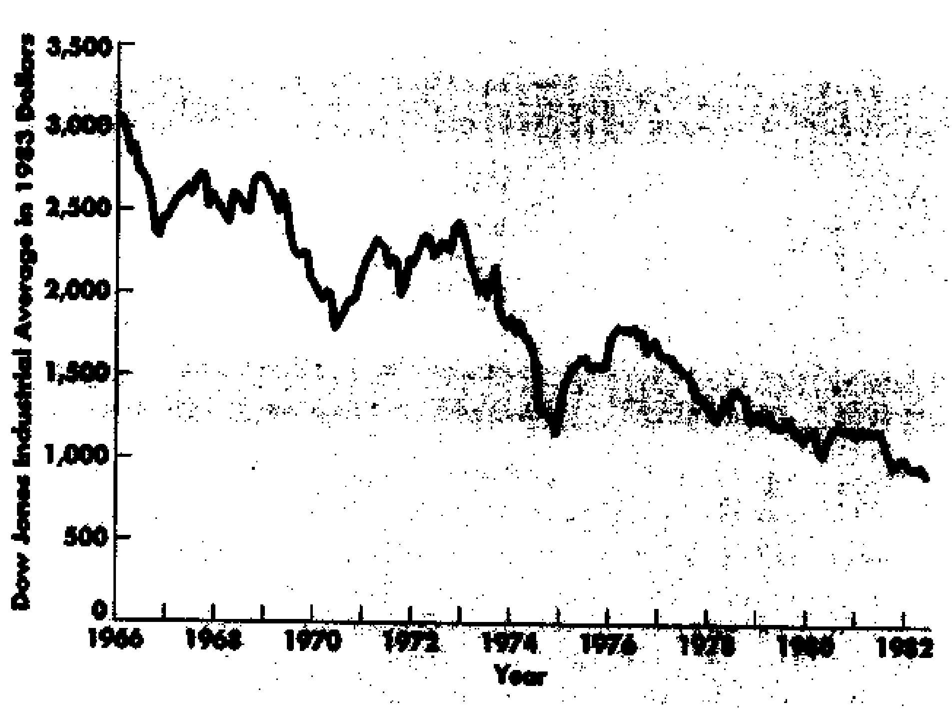
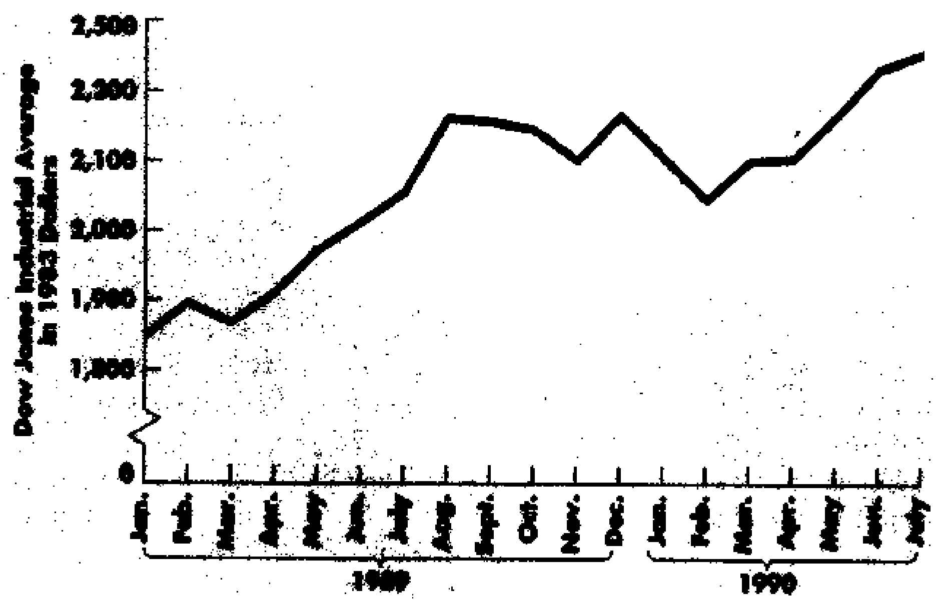


FIGURE 2-15 STOCK PRICES, JANUARY 1989–JULY 1990

This graph seems to indicate that the value of stocks is steadily climbing.

SOURCE: *Economic Report of the President 1991*, Washington, D.C.: U.S. Government Printing Office, 1991.



low (although this may be unavoidable if the graph simply ends with the most recent data).

3. Make sure that (in the absence of some special justification) the graph does not depict only a very brief period of time, which can easily be atypical—like Figure 2-15.

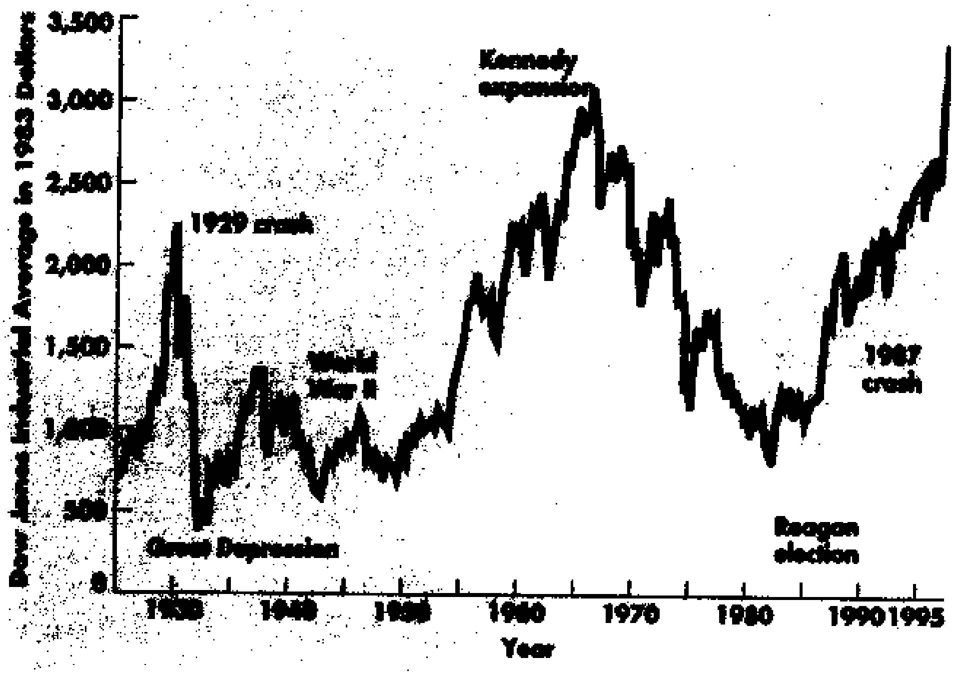
DANGERS OF OMITTING THE ORIGIN

Frequently, the value of an economic variable described by a graph does not fall anywhere near zero during the period under consideration. For example, between 1988 and 1992 the civilian unemployment rate never fell below 5 percent. This

FIGURE 2-16 FULL HISTORY OF STOCK PRICES, CORRECTED FOR INFLATION, 1925–1995

Here we see that stock prices have lots of *both* ups and downs, and that they have not risen nearly as much, over three quarters of a century, as is popularly supposed—after they have been corrected for the fall in the purchasing power of the dollar that resulted from inflation.

SOURCE: *Economic Report of the President 1996*, Washington, D.C.: U.S. Government Printing Office, 1996.



means that a time-series graph showing unemployment over time would have a lot of wasted space between the horizontal axis (where the unemployment rate is zero) and the level of the graph representing 5 percent. There are simply no data points to plot in that range. It is therefore tempting to eliminate this wasted space by beginning the graph at 5 percent, as the magazine *The International Economy* did when it published Figure 2-17 in 1992.

What is wrong with the drawing? Nothing, if you read carefully. But a hasty glance would vastly exaggerate the rise in unemployment. Figure 2-17 makes it look like the United States experienced an economic catastrophe from 1990 to 1992, with unemployment exploding. A less misleading graph, which includes the origin as well as all the "wasted space" in between, is shown in Figure 2-18. Note how this alternative presentation gives a dramatically different visual impression.

Omitting the origin in a graph is dangerous because it exaggerates the magnitudes of the changes that have taken place.

Sometimes, it is true, the inclusion of the origin would waste so much space that it is undesirable to include it. In that case, a good practice is to put a clear warning on the graph to remind the reader that this has been done. Figure 2-19 shows one way to do so.

UNRELIABILITY OF STEEPNESS AND CHOICE OF UNITS

The last pitfall we will consider has consequences similar to the one just discussed. The problem is that we can never trust the visual impression we get from the steepness of a graph. A graph of stock market prices that moves uphill sharply (has a large positive slope) appears to suggest that prices are rising rapidly, while another graph that climbs more slowly seems to imply that prices are going up sluggishly. Yet, depending on how one draws the graph, exactly the same statistics can produce a graph that is rising quickly or slowly.

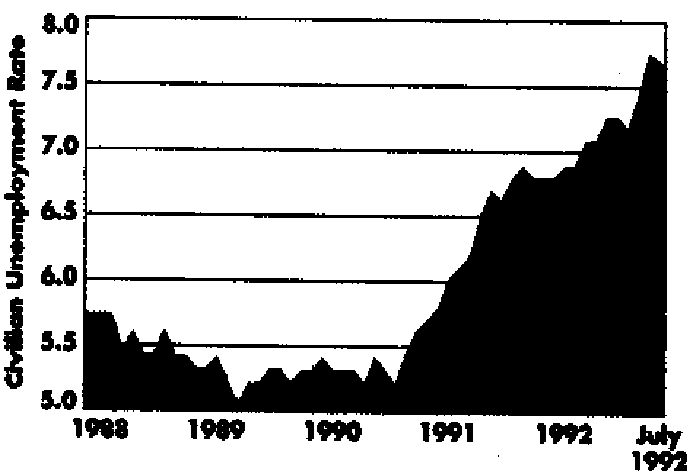
The reason for this possibility is that, in economics, there are no fixed units of measurement. Coal production can be measured in hundredweights (hundreds of pounds) or in tons. Prices can be measured in cents or dollars or millions of dollars. Time can be measured in days or months or years. Any one of these choices is perfectly legitimate, but it makes all the difference to the speed with which a graph using the resulting figures rises or falls.

An example will bring out the point. Suppose we have the following imaginary figures on daily production from a coal mine, which we measure both in hundredweights and in tons (remembering that 1 ton = 20 hundredweights):

Year	Production in Tons	Production in Hundredweights
1980	5,000	100,000
1985	5,050	101,000
1990	5,090	101,800

Look at Figures 2-20(a) and 2-20(b), one graph showing the figures in tons and the other showing the figures in hundredweights. The line looks quite flat in one panel, but quite steep in the other.

FIGURE 2-17 A GRAPH DISTORTED BY OMISSION OF THE ORIGIN



A hasty glance at this figure seems to show that, from mid-1990 to mid-1992, unemployment in the United States soared to disastrous heights.

SOURCE: *The International Economy*, September/October 1992, p. 10.

FIGURE 2-18 THE SAME GRAPH AS FIGURE 2-17, WITH ORIGIN POINT INCLUDED

Extending the previous graph all the way to zero unemployment shows that the rise in unemployment, while significant, was not so enormous as the

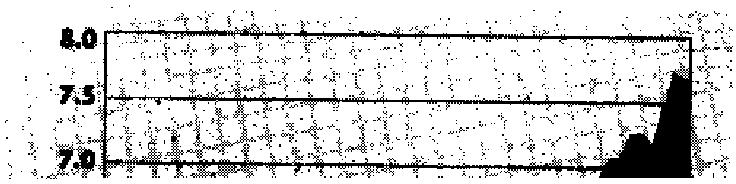
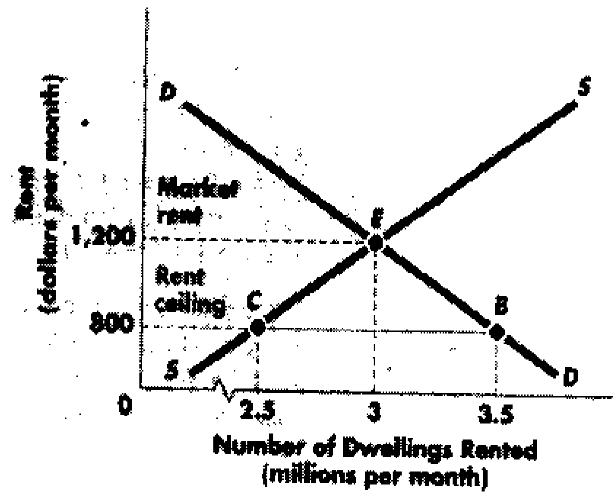


FIGURE 4-10 SUPPLY-DEMAND DIAGRAM FOR RENTAL HOUSING

When market forces are permitted to set rents, the quantity of dwellings supplied will equal the quantity demanded. But when a rent ceiling forces rent below the market level, the number of dwellings supplied (point C) will be less than the number demanded (point B). Thus, rent ceilings induce housing shortages.



duction rather than grow crops at a loss. The result has been thousands of lost jobs and widespread food shortages.

A CASE STUDY: RENT CONTROLS IN NEW YORK CITY

These points and others are best illustrated by considering a concrete example of price ceilings. New York is the only major city in the United States that has had rent controls continuously since World War II. The objective of rent control is, of course, to protect the consumer from high rents. But most economists believe that rent control does not help the cities or their residents and that, in the long run, it makes almost everyone worse off. Elementary supply-demand analysis shows us why.

Figure 4-10 is a supply-demand diagram for rental units in New York. Curve *DD* is the demand curve and the curve *SS* is the supply curve. Without controls, equilibrium would be at point *E*, where rents average $1,200 per month and 3.0 million units are occupied. If rent controls are effective, they must set a ceiling price *below* the equilibrium price of $1,200. But with a low rent ceiling, such as $800, the quantity of housing demanded will be 3.5 million units (point *B*) while the quantity supplied will be only 2.5 million units (point *C*).

The diagram shows a shortage of 1 million apartments. This theoretical concept of a "shortage" manifests itself in New York City as an abnormally low vacancy rate—typically about half of the national urban average. Naturally, rent controls have spawned a lively black market in New York. The black market raises the effective price of rent-controlled apartments in many ways, including bribes, "key money" paid to move up on the waiting list, and requirements that force prospective tenants to purchase worthless furniture at inflated prices.

According to the diagram, rent controls reduce the quantity supplied from 3.0 million to 2.5 million apartments. How does this show up in New York? First, some property owners, discouraged by the low rents, have converted apartment buildings into office space or other uses. Second, some apartments have been inadequately maintained. After all, rent controls create a shortage which makes even dilapidated apartments easy to rent. Third, some landlords have actually abandoned their buildings rather than pay rising tax and fuel bills. These abandoned buildings rapidly become eyesores and eventually pose threats to public health and safety.

year.

With all of these problems, why do rent controls persist in New York City? And why do other cities sometimes move in the same direction?

Part of the explanation is that most people simply do not understand the problems that rent controls create. Another part is that landlords are unpopular politically. But a third, and important, part of the explanation is that not everyone is hurt by rent controls—and those who benefit from controls fight hard to preserve them. In New York, for example, many tenants pay rents that are only a fraction of what their apartments would fetch on the open market. They are, naturally enough, quite happy with this situation. This last point illustrates another very general phenomenon:

Virtually every price ceiling or floor creates a class of people who benefit from the regulations. These people use their political influence to protect their gains by preserving that *status quo*, which is one reason why it is so hard to eliminate price ceilings or floors.

RESTRAINING THE MARKET MECHANISM: PRICE FLOORS

A **PRICE FLOOR** is a legal minimum on the price that may be charged for a commodity.

Interferences with the market mechanism are not always designed to keep prices *low*. Agricultural price supports and minimum wages are two notable examples in which the law keeps prices *above* free-market levels. **Price floors** are typically accompanied by a standard set of symptoms:

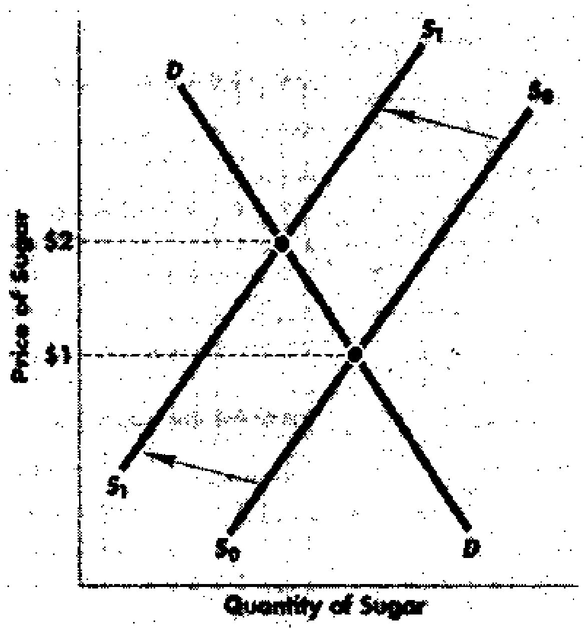
1. *A surplus develops as sellers cannot find enough buyers. Example:* Surpluses of various agricultural products have been a persistent—and costly—problem for the U.S. government. The problem is even worse in the European Union, where the so-called *common agricultural policy* holds prices even higher. Fortunately, agricultural price supports are scheduled to come down under the world trade agreement that took effect in 1995.
2. *Where goods, rather than services, are involved, the surplus creates a problem of disposal.* Something must be done about the excess of quantity supplied over quantity demanded. *Example:* The U.S. government has often been forced to purchase, store, and then dispose of large amounts of surplus agricultural commodities.
3. To get around the regulations, *sellers may offer discounts in disguised—and often unwanted—forms. Example:* When airline fares were regulated by the government, airlines offered more and better food and stylish uniforms for flight attendants instead of lowering fares. Today, the food is worse, but tickets cost much less.
4. *Regulations that keep prices artificially high encourage overinvestment in the industry.* Even inefficient businesses whose high operating costs would doom them in an unrestricted market can survive beneath the shelter of a generous price floor. *Example:* This is why the airline and trucking industries both went through painful "shake outs" of the weaker companies in the 1980s.

Once again, a specific example is useful.

A CASE STUDY: SUGAR PRICE SUPPORTS

America's extensive program of farm price supports began in 1933 as a "temporary method of dealing with an emergency": Farmers were going broke in droves. It is still with us today, even though the farm population of the United

FIGURE 4-11 SUPPORTING THE PRICE OF SUGAR



By inducing domestic producers to grow less sugar, and by limiting imports, the government shifts the supply curve inward to the right from S_0S_0 to S_1S_1. The predictable result is a higher price in the United States.

States is less than a sixth of what it was then.[6] One of the more controversia farm price supports involves the sugar industry. Sugar producers receive low interest loans from the federal government and a guarantee that the price o sugar will not fall below a certain level.

But in a market economy like ours, Congress cannot simply set prices b decree; it must take some action to enforce the price floor. In the case of sugar that "something" is limiting both domestic production and foreign imports thereby shifting the supply curve inward to the left. Figure 4-11 displays th mechanics. Government policies shift the supply curve inward from S_0S_0 to S_1S and drive the U.S. price up from $1 to $2 per pound. The more the supply curv shifts inward, the higher the price.

The sugar industry obviously benefits from the price-control program. Bu consumers pay for it in the form of higher prices for sugar and sugar-filled prod ucts such as soft drinks, candy bars, and cookies. Estimates vary, but the federa sugar price support program costs consumers somewhere around $1.5 billion a year.

If all of this sounds a bit abstract to you, take a look at the ingredients of a U.S.-made soft drink. Instead of sugar, you will likely find "high-fructose corr syrup" listed as a sweetener. Foreign products generally use sugar. But in the United States, sugar is simply too expensive to be used for this purpose.

A CAN OF WORMS

Our two case studies—rent controls and sugar price supports—illustrate some of the major side effects of price floors and ceilings, but barely hint at others. There are difficulties that we have not even mentioned, for the market

[6]As this edition was being written, Congress was considering major legislation that would change the nature of agricultural price supports.

mechanism is a tough bird that imposes suitable retribution on those who seek to circumvent it by legislative decree. Here is a partial list of other problems that may arise when prices are controlled.

FAVORITISM AND CORRUPTION

When price ceilings or floors create shortages or surpluses, someone must decide who gets to buy or sell the limited quantity that is available. This can lead to discrimination along racial or religious lines, political favoritism, or corruption in government. For example, many prices were held at artificially low levels in the former Soviet Union, making queuing for certain goods quite common. But, somehow, Communist party officials and other favored groups were able to purchase the scarce commodities that others could not get.

UNENFORCEABILITY

Attempts to limit prices are almost certain to fail in industries with numerous suppliers, simply because the regulating agency must monitor the behavior of so many sellers. People will usually find ways to evade or violate the law, and something akin to the free-market price will generally reemerge. But there is an important difference: Since the evasion mechanism, whatever its form, will have some operating costs, those costs must be borne by someone. Normally, that someone is the consumer.

AUXILIARY RESTRICTIONS

Fears that a system of price controls will break down invariably lead to regulations designed to shore up the shaky edifice. Consumers may be told when and from whom they are permitted to buy. The powers of the police and the courts may be used to prevent the entry of new suppliers. Occasionally, an intricate system of market subdivision is imposed, giving each class of firms a protected sphere in which others are not permitted to operate. For example, there are laws banning conversion of rent-controlled apartments to condominiums in New York City.

LIMITATION OF VOLUME OF TRANSACTIONS

To the extent that controls succeed in affecting prices, they can be expected to reduce the volume of transactions. Curiously, this is true regardless of whether the regulated price is *above* or *below* the free-market equilibrium price. If it is set above the equilibrium price, quantity demanded will be below the equilibrium quantity. On the other hand, if the imposed price is set below the free-market level, quantity supplied will be cut down. Since sales volume cannot exceed either the quantity supplied or the quantity demanded, a reduction in the volume of transactions is the result.[7]

MISALLOCATION OF RESOURCES

Departures from free-market prices are likely to result in misuse of the economy's resources because the connection between production costs and prices is broken. For example, Russian farmers used to feed their farm animals bread instead of unprocessed grains because price ceilings kept the price of bread ludicrously low. In addition, just as more complex locks lead to more sophisticated

---

[7]See Review Question 9 at the end of the chapter.

burglary tools, more complex regulations lead to the use of yet more resources for their avoidance.

Economists put it this way. Free markets are capable of dealing with the three basic coordination tasks outlined in Chapter 3: deciding *what* to produce, *how* to produce it, and *to whom* the goods should be distributed. Price controls throw a monkey wrench into the market mechanism. Though the market is surely not flawless, and government interferences often have praiseworthy goals, good intentions are not enough. Any government that sets out to repair what it sees as a defect in the market mechanism runs the risk of causing even more serious damage elsewhere. As a prominent economist once quipped, societies that are too willing to interfere with the operation of free markets soon find that the invisible hand is nowhere to be seen.

A SIMPLE BUT POWERFUL LESSON

The lessons you have learned in this chapter may seem elementary, even obvious. In many respects, they are. But they are also very important, indeed, indispensable. Although the law of supply and demand is one of the simplest principles in economics, it is also one of the most powerful. Astonishing as it may seem, many people in authority fail to understand the law of supply and demand or cannot apply it to concrete situations.

For example, a few years ago *The New York Times* carried a dramatic front page picture of the president of Kenya setting fire to a large pile of elephant tusks that had been confiscated from poachers. The accompanying story explained that the burning was intended as a symbolic act to persuade the world to halt the ivory trade.[8] Economists claim no expertise on the likely psychological effect of burning elephant tusks, though one may doubt that it touched the hearts of criminal poachers. However, one economic effect was clear. By reducing the supply of ivory on the world market, the burning of tusks forced up the price of ivory, which raised the illicit rewards reaped by those who slaughter elephants. That could only encourage more poaching—precisely the opposite of what the Kenyan government sought to accomplish.

[8] *The New York Times*, July 19, 1989.

SUMMARY

1. The quantity of a product that is demanded is not a fixed number. Rather, **quantity demanded** depends on such factors as the price of the product, consumer incomes, and the prices of other products.
2. The relationship between quantity demanded and price, holding all other things constant, can be displayed graphically on a **demand curve.**
3. For most products, the higher the price, the lower the quantity demanded. So the demand curve usually has a negative slope.
4. The quantity of a product that is supplied also depends on its price and many other influences. A **supply curve** is a graphical representation of the relationship between **quantity supplied** and price, holding all other influences constant.
5. For most products, the supply curves have positive slopes, meaning that higher prices call forth greater quantities supplied.
6. A market is said to be in **equilibrium** when quantity supplied is equal to quantity demanded. The equilibrium price and quantity are shown by the point on a graph where the supply and demand curves intersect. The **law of supply and demand** states that price and quantity tend to gravitate to this point in a free market.

7. A change in quantity demanded that is caused by a change in the price of the good is represented by a *movement along a fixed demand curve*. A change in quantity demanded that is caused by a change in any other determinant of quantity demanded is represented by a *shift of the demand curve*.
8. This same distinction applies to the supply curve: Changes in price lead to *movements along a fixed supply curve;* changes in other determinants of quantity supplied lead to *shifts of the whole supply curve*.
9. Changes in consumer incomes, tastes, technology, prices of competing products, and many other influences cause shifts in either the demand curve or the supply curve and produce changes in price and quantity that can be determined from **supply-demand diagrams.**
10. An attempt by government regulations to force prices above or below their equilibrium levels is likely to lead to **shortages or surpluses,** to black markets in which goods are sold at illegal prices, and to a variety of other problems. This is one of the **Ideas for Beyond the Final Exam.**

KEY TERMS

Quantity demanded
Demand schedule
Demand curve
Quantity supplied
Supply schedule
Supply curve
Supply-demand diagram
Shortage
Surplus
Equilibrium price and quantity
Equilibrium
Law of supply and demand
Shifts in versus movements along supply and demand curves
Price ceiling
Price floor

QUESTIONS FOR REVIEW

1. How often do you rent videos? Would you do so more often if a rental cost half as much? Distinguish between your demand curve for home videos and your "quantity demanded" at the current price.
2. What shapes would you expect for demand curves for the following:
   a. A medicine that means life or death for a patient?
   b. French fries in a food court with stands offering many types of food?
3. The following are the assumed supply and demand schedules for hamburgers in Collegetown:

Demand Schedule		Supply Schedule	
Price	Quantity Demanded (per year)	Price	Quantity Supplied (per year)
$2.00	6,000	$2.00	15,000
1.75	8,000	1.75	14,000
1.50	10,000	1.50	13,000
1.25	12,000	1.25	12,000
1.00	14,000	1.00	11,000
0.75	16,000	0.75	10,000

   a. Plot the supply and demand curves and indicate the equilibrium price and quantity.
   b. What effect would a decrease in the price of beef (a factor of production) have on the equilibrium price and quantity of hamburgers, assuming all other things were to remain constant? Explain your answer with the help of a diagram.
   c. What effect would an increase in the price of pizza (a substitute commodity) have on the equilibrium price and quantity of hamburgers, assuming again that all other things were held constant? Use a diagram in your answer.
4. Suppose the supply and demand schedules for bicycles are as follows:

Price	Quantity Demanded (per year)	Quantity Supplied (per year)
$ 80	20	12
100	18	14
120	16	16
140	14	18
160	12	20
180	10	22

   a. Graph these curves and show the equilibrium price and quantity.
   b. Now suppose that it becomes unfashionable to ride a bicycle, so the quantity demanded at each price falls by 4 million bikes per year. What is the new equilibrium price and quantity? Show this solution graphically. Explain why the quantity falls by less than 4 million bikes per year.

c. Suppose *instead* that several major bicycle producers go out of business, thereby reducing the quantity supplied by 4 million bikes at every price. Find the new equilibrium price and quantity, and show it graphically. Explain again why quantity falls by less than 4 million.

d. What are the equilibrium price and quantity if the shifts described in Review Questions 4(b) and 4(c) happen at the same time?

5. The table below summarizes information about the market for principles of economics textbooks:

Price	Quantity Demanded (per year)	Quantity Supplied (per year)
$30	2,100	100
40	1,100	300
50	600	600
60	350	1,000
70	225	1,500

a. What is the market equilibrium price and quantity of textbooks?

b. In order to quell outrage over tuition increases, the college places a $40 limit on the price of textbooks. How many textbooks will be sold now?

c. While the price limit is still in effect, automated publishing increases the efficiency of textbook production. Show graphically the likely effect of this innovation on the market price and quantity.

6. Show how the following demand curves are likely to shift in response to the indicated changes:

a. The effect of a drought on the demand curve for umbrellas.

b. The effect of higher popcorn prices on the demand curve for movie tickets.

c. The effect on the demand curve for coffee of a decline in the price of Dr Pepper soda.

7. Discuss the likely effects of:

a. Rent ceilings on the market for apartments.

b. Floors under wheat prices on the market for wheat.

Use supply-demand diagrams to show what may happen in each case.

8. U.S. government price supports for milk lead to a chronic surplus of milk. In an effort to reduce the surplus about a decade ago, Congress offered to pay dairy farmers to slaughter cows. Use two diagrams, one for the milk market and one for the meat market, to illustrate how this policy should have affected the price of meat. (Assume that meat is sold in an unregulated market.)

9. On page 88, it is claimed that either price floors or price ceilings reduce the actual quantity exchanged in a market. Use a diagram or diagrams to support this conclusion, and explain the common sense behind it.

10. The same rightward shift of the demand curve may produce a very small or a very large increase in quantity, depending on the slope of the supply curve. Explain with diagrams.

11. In 1981, when regulations were holding the price of natural gas below its free-market level, then-Congressman Jack Kemp of New York said the following in an interview with *The New York Times*: "We need to decontrol natural gas, and get production of natural gas up to a higher level so we can bring down the price."[9] Evaluate the congressman's statement.

12. From 1979 to 1989 in the United States, the number of working men grew 12 percent while the number of working women grew 29 percent. During this time, average wages for men fell slightly while average wages for women rose about 7 percent. Which of the following two explanations seems most consistent with the data?

a. Women decided to work more, raising their relative supply (relative to men).

b. Discrimination against women declined, raising the relative (to men) demand for female workers.

13. The two diagrams on the next page show supply and demand curves for two substitute commodities: tapes and compact discs (CDs).

a. On the right-hand diagram, show what happens when rising raw material prices make it costlier to produce tapes.

b. On the left-hand diagram, show what happens to the market for CDs.

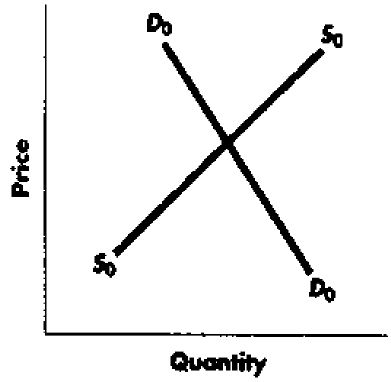
14. (More difficult) Consider the market for milk discussed in this chapter (Tables 4-1 through 4-3 and Figures 4-1 through 4-3). Suppose that the government decides to fight kidney stones by levying a tax of 30 cents per quart on sales of milk. Follow these steps to analyze the effects of the tax:

a. Construct the new supply schedule (to replace Table 4-2) that relates quantity supplied to the price that consumers pay. (*Hint:* Before the tax, when consumers paid 70 cents, farmers supplied 60 billion quarts. With a 30 cent tax, when consumers pay 70 cents, farmers will receive only 40 cents. Table 4-2 tells us that they will provide only 30 billion quarts at this price. This is one point on the new supply curve. The rest of the curve can be constructed in the same way.)

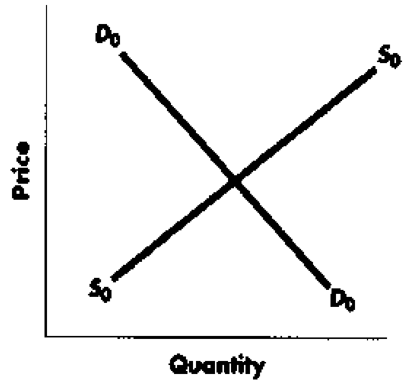
b. Graph the new supply curve constructed in Review Question 14(a) on the supply-demand diagram depicted in Figure 4-3. What are the new equilibrium price and quantity?

---

[9] *The New York Times*, December 23, 1981.



Compact Discs
(a)



Tapes
(b)

c. Does the tax succeed in its goal of reducing the consumption of milk?

d. How much does the equilibrium price increase? Is the price rise greater than, equal to, or less than the 30 cent tax?

e. Who actually pays the tax, consumers or producers? (This may be a good question to discuss in class.)

15. (More difficult) The demand and supply curves for T-shirts in Touristown, U.S.A., are given by the following equations:

$$Q = 24{,}000 - 500P \quad Q = 6{,}000 + 1{,}000P$$

where P is measured in dollars and Q is the number of T-shirts per year.

a. Find the equilibrium price and quantity algebraically.

b. If tourists decide they do not really like T-shirts that much, which of the following might be the new demand curve?

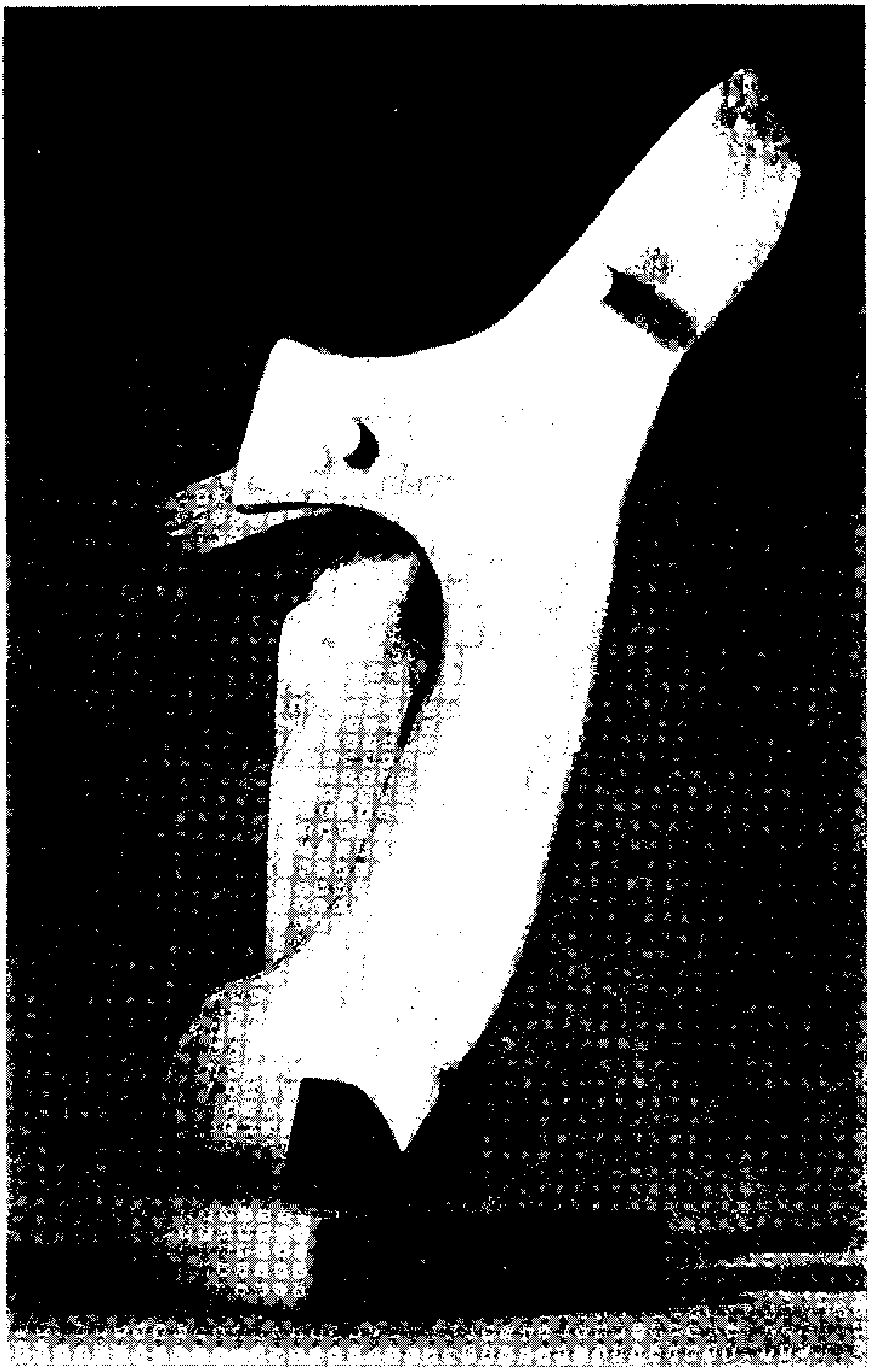
$$Q = 21{,}000 - 500P \quad Q = 27{,}000 - 500P$$

Find the equilibrium price and quantity after the shift of the demand curve.

c. If, *instead*, two new stores that sell T-shirts open up in town, which of the following might be the new supply curve?

$$Q = 3{,}000 + 1{,}000P \quad Q = 9{,}000 + 1{,}000P$$

Find the equilibrium price and quantity after the shift of the supply curve.



PART II

ESSENTIALS OF MICROECONOMICS: CONSUMERS AND FIRMS

CHAPTER 5

CONSUMER CHOICE: THE DEMAND SIDE OF THE MARKET

Everything is worth what its purchaser will pay for it.
Publilius Syrus (1st Century B.C.)

This chapter and the three that follow describe and analyze the basic building blocks with which economists explain how markets work. The two essential elements of any market are its buyers (consumers) and its sellers (producers). As in an automobile engine, all the parts of a market function together and simultaneously, so there is no one logical place to begin the story. Furthermore, the heart of the story is not found in the individual components, but in the way they fit together.

Still, we must begin somewhere, so we will follow custom and start off with the consumer. Accordingly, this chapter discusses the logic of the choices of individual consumers and how those choices lead to demand curves for individual consumers. Then, at the end of the chapter and in Chapter 6, we bring together all the consumers in a given market and examine the market demand curve.

Since an individual consumer's demand curve indicates how much of a good he or she wants to buy at each possible price, its explanation must, ultimately, be a matter of consumer psychology. But economists claim no special qualifications for making deep pronouncements about consumer psychology, so our exploration will not go very far below the surface. We will, however, describe methods by which economists do analyze consumer choice, and we will cast some light on a number of important issues, including the consequences of scarcity for consumer behavior.

In so doing, this chapter will also introduce you to one of the most powerful and widely used of all the economist's tools: *marginal analysis*. The marginalist way of thinking provides invaluable guidance any time a decision maker confronts a range of choices. How many compact discs (CDs) should a consumer purchase? How many copies of a particular CD should the manufacturer produce? What price should the manufacturer charge for each copy? You will see that marginal analysis offers a guide to the *optimal decision* in these and many other cases.

Once we complete all of this study, we will turn to the supply side of the market, beginning with analysis in Chapter 7 of production decisions—how business firms decide what inputs to use to produce the goods and services that they supply to consumers. Chapter 8 completes the basic analysis of the market mechanism by investigating how outputs and prices are determined.

Armed with this analysis of the market mechanism, we will go on to cover specific applications in subsequent chapters. We will study, in particular, what the market does well and what its failings are. We will examine competitive and monopolized markets and markets that lie in between the two. We will also study what this analysis suggests about the appropriate role of government in cases where the market performs imperfectly.

A PARADOX: SHOULD WATER BE WORTH MORE THAN DIAMONDS?

When Adam Smith was lecturing at the University of Glasgow in the 1760s, he introduced the study of demand by posing a puzzle. Common sense, he said, suggests that the price of a commodity must somehow depend on what that good is worth to consumers—on the amount of *utility* that the commodity offers. Yet, Smith pointed out, some cases suggest that a good's utility may have little influence on its price.

Smith cited diamonds and water as examples. He noted that water has enormous value to most consumers; indeed, its availability can be a matter of life and death. Yet water generally is either free or sells at a very low price, while diamonds sell for very high prices even though they hardly constitute necessities. In a few pages, we will be in a position to see how marginal analysis helps to resolve the paradox.

SCARCITY AND DEMAND

The choices of any individual consumer are subject to one overriding constraint that is at least partly beyond the consumer's control: the individual has only a limited income available to spend. This scarcity of income is the obvious reason why less affluent consumers demand fewer computers, trips to foreign countries, and expensive restaurant meals than wealthy consumers.

Because income is limited (a scarce resource) purchase decisions for the different commodities bought by any consumer must be *interdependent*. The number of movies that Jane can afford to see depends on the amount she spends on new clothing. If Sam's parents have just sunk a lot of money into an expensive addition to their home, they may have to give up a vacation trip. Thus, no one can truly understand the demand curves for movies and clothing, or for homes and vacation trips, without considering them together.

The quantity of movies demanded, for example, probably depends not only on ticket prices but also on the prices of clothing. Thus, a big sale on jeans might induce Jane to splurge on several pairs, leaving her with little or no cash to spend on movies. Therefore, an analysis of consumer demand that focuses on only one

commodity at a time leaves out an essential part of the story. Nevertheless, to make the analysis easier to follow, we will begin by considering products in isolation. Later in the chapter and in the appendix, we will tell the fuller story.

TOTAL AND MARGINAL UTILITY

In the American economy, millions of consumers make millions of decisions every day. You decide to buy a movie ticket instead of a paperback novel. Your roommate decides to buy 2 pounds of imported cheese rather than 1 pound or 3 pounds. How do people make these decisions?

Economists have constructed a simple theory of consumer choice based on the hypothesis that each consumer spends his or her income in the way that yields the greatest amount of satisfaction, or *utility*. This seems to be a reasonable starting point, since it says little more than that people do what they prefer. To make the theory operational, we need a way to measure utility.

A century ago, economists envisioned utility as an indicator of the pleasure a person derives from consuming some set of goods, and they thought that utility could be measured directly in some kind of psychological units (sometimes called *utils*) after somehow reading the consumer's mind. Gradually, they came to realize that this was an unnecessary and, perhaps, impossible task. How many utils did you get from the last movie you saw? You probably cannot answer that question because you have no idea what a util is. Neither does anyone else.

But you may be able to answer a different question like, "How many hamburgers would you give up to get that movie ticket?" If you answer "three," no one can say how many utils you get from seeing a film, but they can say that you get more from the movie than from a single hamburger. When economists approach the issue in this manner, hamburgers, rather than utils, become the unit of measurement. They can say that the utility of a movie (to you) is three hamburgers.

Early in the 20th century, economists concluded that this indirect way of measuring utility gave them all they needed to build a theory of consumer choice. One can measure the utility of a movie ticket by asking how much of some other commodity (like hamburgers) you are willing to give up for it. Any commodity will do for this purpose, but the simplest choice, and the one that we will use in this book, is money.[1]

The **TOTAL UTILITY** of a quantity of a good to a consumer (measured in money terms) is the maximum amount of money that he or she is willing to give in exchange for it.

Thus, we define the **total utility** of some bundle of goods to some consumer as *the largest sum of money that person will voluntarily give up in exchange for those goods*. For example, imagine that Sam Sophomore is considering whether to buy six servings of Chicken Maknoogats during the next month. He loves Maknoogats, but he has already paid for the terrible food at the college cafeteria. After careful figuring, Sam decides that he will not buy the six servings if they cost more than $22.20 in total, but that he will buy them if they cost $22.20 or less. Then the *total utility* of six portions of Chicken Maknoogats to him is $22.20—the maximum amount he is willing to spend to have them.

[1]NOTE TO INSTRUCTORS: You will recognize that, while not using the terms, we are distinguishing here between *neoclassical cardinal utility* and *ordinal utility*. Moreover, throughout the book, *marginal utility in money terms* (or *money marginal utility*) is simply used as a synonym for the *marginal rate of substitution* between money and the commodity.

The **MARGINAL UTILITY** of a commodity to a consumer (measured in money terms) is the maximum amount of money that he or she is willing to pay *for one more unit* of that commodity.

Total utility measures the benefit that Sam derives from his purchases. It is *total* utility that really matters to Sam. But to understand which decisions most effectively promote total utility, we must consider the related concept of **marginal utility.** This term refers to *the addition to total utility that an individual derives by consuming one more unit of any good.* That is, if Sam consumed five servings of Maknoogats last month, marginal utility indicates how much additional pleasure he would have received by increasing his consumption to six servings instead.

Table 5-1 helps to clarify the distinction between marginal and total utility and shows how the two are related. The first two columns show how much *total* utility (measured in money terms) Sam derives from various quantities of Chicken Maknoogats, ranging from zero to eight servings per month. For example, a single portion is worth (no more than) $6 to him, two servings are worth $11.60, and so on. The *marginal* utility is the *difference* between any two successive total utility figures. For example, if Sam already has consumed three servings (worth $16.00 to him), an additional serving brings his total utility up to $19.60. His marginal utility is thus the difference between the two, or $3.60.

Remember: Whenever we use the terms *total utility* and *marginal utility,* we define them in terms of the consumer's willingness to part with money for the commodity—not in some unobservable (and imaginary) psychological units.

THE "LAW" OF DIMINISHING MARGINAL UTILITY

With these definitions, we can now propose a simple hypothesis about consumer tastes: The more of a good a consumer has, the less *marginal* utility an additional unit has.

In general, this is a plausible proposition, and it is widely used in economics. The idea is based on the assertion that every person has a hierarchy of uses for a particular commodity. All of these uses are valuable, but some are more valuable than others. Let's consider Chicken Maknoogats again. Instead of eating all the Maknoogats he buys, Sam may decide to throw a Maknoogats party. The first person he will invite, of course, is his girlfriend. If he feels that he can afford more guests, he may also invite a roommate; if he feels really flush, he may even

TABLE 5-1

TOTAL AND MARGINAL UTILITY OF CHICKEN MAKNOOGATS (MEASURED IN MONEY TERMS)

(1) Quantity (Q) (servings per month)	(2) Total Utility (TU) (in dollars)	(3) Marginal Utility (MU) = (ΔTU/ΔQ) (in dollars)	(4) Point in Figure 5-1
0	0.00		A
		6.00	
1	6.00		B
		5.60	
2	11.60		C
		4.40	
3	16.00		D
		3.60	
4	19.60		E
		1.80	
5	21.40		F
		0.80	
6	22.20		G
		0.40	
7	22.60		H
		0.00	
8	22.60		

Note: Each entry in Column (3) is the difference between successive entries in Column (2).

invite an instructor. Thus, if he buys only one serving per month, he will eat it himself. If he buys a second serving, he will share it with his girlfriend. He may share a third with his roommate. If he decides he can afford a fourth serving, he may invite his instructor to join the party.

The point is obvious. Each serving contributes something to Sam's satisfaction. But each *additional* serving contributes less (measured in terms of money) than its predecessor because its use has a lower priority. This, in essence, is the logic behind the "law" **of diminishing marginal utility.**

The **"LAW" OF DIMINISHING MARGINAL UTILITY** asserts that additional units of a commodity are worth less and less to a consumer in money terms. As the individual's consumption increases, the marginal utility of each additional unit declines.

The third column of Table 5-1 illustrates this concept. The marginal utility (abbreviated MU) of the first serving of Chicken Maknoogats is $6; that is, Sam is willing to pay *up to* $6 for the first serving. The second serving is worth no more than $5.60, the third serving only $4.40, and so on until Sam is willing to pay only $0.80 for the sixth serving (the MU of that serving is $0.80).

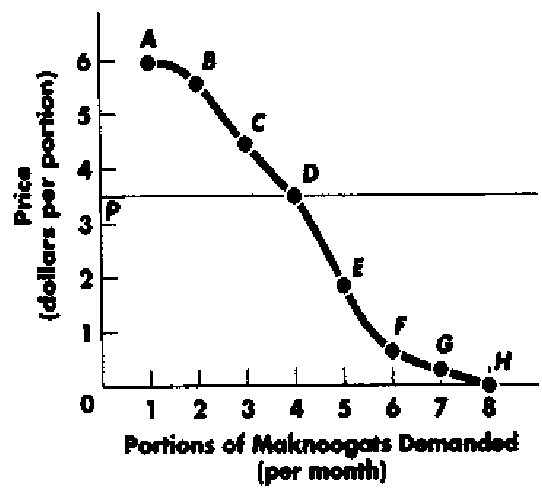
Figure 5-1 shows a graph of the numbers in the first and third columns of the table. For example, point *D* indicates that the MU of a fourth serving is $3.50. Note that the curve for marginal utility is negatively sloped; this again illustrates how marginal utility diminishes as the quantity of the good rises.

Like most laws, however, it has exceptions. Some people want even more of some good that is particularly significant to them as they acquire more. The needs of alcoholics and stamp collectors are good examples. The stamp collector who has a few stamps may consider the acquisition of one more to be mildly amusing. The person who has a large and valuable collection may be prepared to go to the ends of the earth for another stamp. Similarly, the alcoholic who finds a first dry martini quite pleasant may find the fourth or fifth to be absolutely irresistible. Economists generally treat such cases of increasing marginal utility as anomalies. For most goods and most people, marginal utility declines as consumption increases.

Table 5-1 and Figure 5-1 illustrate another noteworthy relationship. Notice that as someone buys more and more units of the commodity, that is, as that person moves further down the table, the total utility numbers get larger and larger, while the marginal utility numbers get smaller and smaller. The reasons should now be fairly clear. The marginal utility numbers keep declining as a result of

FIGURE 5-1 A TYPICAL MARGINAL UTILITY OR DEMAND CURVE. SAM'S DEMAND FOR CHICKEN MAKNOOGATS

This demand curve is derived from the consumer's table of marginal utilities by following the optimal purchase rule. The points in the graph correspond to the numbers in Table 5-1.



the law of diminishing marginal utility, as we just explained. But *total* utility keeps rising so long as marginal utility remains positive. A woman who owns ten golf balls, other things being equal, is better off (she has higher total utility) than a woman who possesses only nine, as long as the MU of the tenth ball is positive. In summary:

As a rule, as a person acquires more of a commodity, total utility increases and marginal utility from that good decreases, all other things being equal. In particular, when a commodity is very scarce, economists expect it to have a high marginal utility, even though it may provide little total utility.

THE OPTIMAL PURCHASE RULE

Now let us put the concept of marginal utility to work in analyzing consumer choice. As already noted, every consumer has a limited amount of money to spend. Consumers must always make choices among the many commodities that compete for their dollars. Which items will they buy and in what quantities?

Sam can obviously buy many different quantities of Maknoogats that will add to his total utility. But which of these quantities will yield him the greatest net benefits? If Maknoogats were all that Sam ever considered buying, in theory the choice would involve a simple calculation. One would need a statistical table that listed all of the alternative numbers of portions that Sam might buy. The table should indicate the net utility for each of these possible choices, that is, the total utility that Sam would get from that number of Maknoogat portions, minus the utility he would lose by having to pay for them. One could then simply read Sam's optimal choice from this imaginary table—the number of portions that would give Sam the highest net utility number.

Even in theory, calculation of the optimal decision is, unfortunately, more difficult. No such table of net utilities exists, an increase in the expenditure on Maknoogats would mean less money available for clothing or movies, and Sam must balance the benefits of spending on each of these items against spending on the others. All of this means that we must find a more effective technique to determine the optimal purchase of Maknoogats (as well as clothing, entertainment, and so on). That technique is *marginal analysis.*

To see how marginal analysis helps to determine the optimal purchase decision for Sam, first recall that he is assumed to be maximizing the total *net* utility he obtains from his Maknoogats purchase. That is, he is trying to select the number of portions that maximizes the total utility the Maknoogats provide him *minus the total utility he gives up because of the money he must pay for them.*

Now, it is useful to compare the analysis of the optimal decision process to the process of climbing a hill. First, Sam considers the possibility of buying only one portion of Maknoogats. Then, instead, he considers buying two portions, and so on. If two portions give him a higher total net utility than one, we may think of him as moving higher up the total net utility hill. Larger portions enable him to ascend that hill higher and higher, until at some quantity he reaches the top—*the optimal purchase quantity.* Then, if he buys any more he will have overshot the peak and begun to descend the hill.

How does marginal analysis help Sam to find that optimal purchase quantity, and how does it warn him if he is planning to purchase too little (so that he is still on the ascending portion of the hill) or too much (so that he is descending)? The numerical example will help reveal the answers. The marginal utility of, for

DO CONSUMERS REALLY BEHAVE "RATIONALLY" AND MAXIMIZE UTILITY?

It may strike you that this chapter's discussion of the consumer's decision process—equating price and marginal utility—does not resemble the thought processes of any consumer you have ever met. Buyers may seem to make decisions much more instinctively and without any formal calculation of marginal utilities or anything like them. That is true, and yet it need not undermine the pertinence of the discussion.

When you give a command to your computer, you actually activate some electronic switches and start some operations in what is referred to as *binary code*. Most computer users do not know they are doing this and do not care. Yet they are doing it nevertheless, and the analysis of the computation process does not misrepresent the facts by describing this sequence. In the same way, if the shopper divides his purchasing power among various available purchase options in a way that yields the largest possible utility for his money, he *must* be following the rules of marginal analysis, even though he is totally unaware of this choice.

Still, growing experimental evidence points out some persistent deviations between reality and the picture of consumer behavior provided by marginal analysis. Recent experimental studies by groups of economists and psychologists have turned up many examples of behavior that seem to violate the optimal purchase rule. For example, one such study offered two groups of respondents what were really identical options, presumably yielding similar marginal utilities. Yet, depending on differences in some irrelevant information that was also provided to the respondents, the two groups made very different choices.

> ... One group of subjects received the information in parentheses, and the other received the information in brackets....
>
> *[Problem 1].* Imagine that you are about to purchase ... a calculator for ($15)[$125]. The calculator salesman informs you that the calculator you wish to buy is on sale for ($10)[$120] at the other branch of the store, located a 20-minute drive away. Would you make the trip to the other store?
>
> ... The responses to the two versions of this problem were quite different. When the calculator cost $125 only 29 percent of the subjects said they would make the trip, whereas 68 percent said they would go when the calculator cost only $15.

Thus, in [this] problem the two groups were told they could save $5 on the price of a product if they took a 20-minute trip to another store. Yet, depending on whether the product was a cheap or an expensive one, the number of persons willing to make the trip to save the same amount of money was very different.

SOURCE: Richard H. Thaler, *Quasi Rational Economics* (New York: Russell Sage Foundation, 1992), pp. 148–150.



example, a third portion of Chicken Maknoogats is $4.40, according to Table 5-1. This means that the total utility Sam obtains from three portions ($16.00) is exactly $4.40 higher than the total utility he gets from two portions ($11.60). Thus, the marginal utility of the third portion is the amount that this portion adds to Sam's total utility. As long as marginal utility is a positive number, the more Sam purchases, the more total utility he will get.

That is the benefit side of the purchase to Sam. But such a transaction also has a debit side—the amount he must pay for the purchase. Suppose that the price is $3.60 per portion. Then the net *marginal* utility of the third portion is marginal utility minus price, $4.40 minus 3.60, or $0.80. This is the amount that the third portion of Maknoogats adds to Sam's total net utility. (See the third and fourth lines of Table 5-1.) So Sam is really better off with three portions than with two.

We can generalize the logic of the previous paragraph to show how marginal analysis solves the problem of finding the optimal purchase quantity.

RULE 1 If marginal net utility is positive, the consumer must be buying too small a quantity to maximize total net utility. Because marginal utility exceeds the price, one can increase total net utility by buying one more unit of the product. In other words, a positive marginal net utility means that total net utility is still going uphill. The consumer has not yet bought enough to get to the top of the hill.

RULE 2 No purchase quantity for which marginal net utility is a negative number can ever be optimal. In such a case, a buyer can get a higher total net utility by cutting back the purchase quantity. Sam would have climbed too far up the net utility hill, passing the topmost point and beginning to descend.

This leaves only one option. The consumer cannot be at the top of the hill if marginal net utility (MU − *P*) is greater than zero, that is, if MU is greater than *P*. Similarly, the purchase quantity cannot be optimal if marginal net utility at that quantity (MU − *P*) is less than zero, that is, if MU is less than *P*. The purchase quantity can be optimal, giving the consumer the highest possible total net utility, only if:

$$\text{Marginal net utility} = \text{MU} - P = 0, \text{ that is, if MU} = P$$

Consequently, the hypothesis that the consumer chooses in a way that makes the largest net contribution to total utility leads to the following *optimal purchase rule:*

It always pays the consumer to buy more of any commodity whose marginal utility (measured in money) exceeds its price, and less of any commodity whose marginal utility is less than its price. When possible, the consumer should buy the quantity of each good at which price (*P*) and marginal utility (MU) are exactly equal, that is, at which

$$\text{MU} = P,$$

because only these quantities will maximize the *net total utility* that the consumer gains from purchases, given the fact that these decisions must divide available money among all purchases.[2]

Notice that, although the consumer really cares about maximizing total net utility and has no intrinsic interest in marginal utility, we have used marginal analysis as a guide to the optimum purchase quantity. Marginal analysis serves only as an analytic method—as a means to an end. In Chapter 8, after several other applications of marginal analysis, we will generalize the discussion and summarize how to apply it to make optimal decisions in a wide variety of fields besides consumer purchases.

Let us briefly review graphically how the underlying logic of the marginal way of thinking leads to the optimal purchase rule: MU = *P*. Refer back to the graph of marginal utilities of Chicken Maknoogats (Figure 5-1). Suppose that

---

[2]Economists can equate a dollar price with marginal utility only because they measure marginal utility in money terms (or, as they more commonly state, because they deal with the marginal rate of substitution of money for the commodity). If marginal utility were measured in some psychological units not directly translatable into money terms, a comparison of *P* and MU would have no meaning. However, MU could also be measured in terms of any commodity other than money. (Example: How many servings of Maknoogats is Sam willing to trade for an additional ticket to a basketball game?)

MakDoogal's currently sells Maknoogats at a price of $3.60 a serving (line *PP* in the graph).

At this price, five servings (point *E*), for example, is *not* an optimal purchase because the $1.80 marginal utility of the fifth serving is less than its $3.60 price. Sam would be better off with only four servings, since that would save $3.60 with only a $1.80 loss in utility—a net gain of $1.80 ($3.60 − $1.80)—from the decision to buy one fewer serving.

The rule for optimal purchases states that Sam should not buy a quantity at which MU is higher than price (points like *A*, *B*, and *C*) because a larger purchase would make him better off. Similarly, he should not end up at points *E*, *F*, *G*, and *H*, at which MU is below price, because he would be better off buying less. Rather, Sam should buy four servings (point *D*), where *P* = MU. Thus, marginal analysis leads naturally to the rule for optimal purchase quantities.

The decision to purchase a quantity of a good that leaves marginal utility greater than price cannot maximize total net utility, because buying an additional unit would add more to total utility than it would cost. Similarly, it cannot be optimal for the consumer to buy a quantity of a good that leaves marginal utility above price, because then a reduction in the quantity purchased would save more money than it would sacrifice in utility. Consequently, the consumer can maximize total net utility only if the purchase quantity makes marginal utility equal to price.

Note that price is an objective, observable figure determined by the market, while marginal utility is subjective and reflects the tastes of the consumer. Since the consumer lacks the power to influence the price, he or she must adjust purchase quantities to make the marginal utility of each good equal to the price given by the market.

FROM MARGINAL UTILITY TO THE DEMAND CURVE

We can use the optimal purchase rule to show that the "law" of diminishing marginal utility implies that demand curves typically slope downward to the right, that is, they have negative slopes.[3] To do this, we use the list of marginal utilities in Table 5-1 to determine how many orders of Chicken Maknoogats Sam would buy at any particular price. For example, we see that at a price of 40 cents, it pays for Sam to buy seven orders, because the MU of the seventh order is $0.40. Table 5-2 gives several alternative prices and the optimal purchase quantity corresponding to each price derived in the way just described. (To make sure you understand the logic behind the optimal purchase rule, verify that the entries in the right-hand column of Table 5-2 are in fact correct.) This *demand schedule* is depicted graphically as Sam's *demand curve* shown in Figure 5-1. This demand curve is simply the orange marginal utility curve. You can see that it has the characteristic negative slope of demand curves.

Let us examine the logic underlying the negatively sloped demand curve a bit more carefully. If Sam is purchasing the optimal number of Maknoogats, and then the price falls, he will find that his marginal utility of that product is now *above* the reduced price. For example, Table 5-1 indicates that at a price of $4.40

[3] If you need to review the concept of slope, turn back to the appendix to Chapter 1.

TABLE 5-2

LIST OF OPTIMAL QUANTITIES OF CHICKEN MAKNOOGATS FOR SAM TO PURCHASE AT ALTERNATIVE PRICES

Price*	Quantity of Purchase (portions per month)
$0.40	7
0.80	6
1.80	5
3.60	4
4.40	3
5.60	2
6.00	1

*Note that for simplicity of explanation the prices shown have been chosen to equal the marginal utilities in Table 5-1. In-between prices would make the optimal choices involve fractions of a portion (say, 2.6 portions).

per serving, it is optimal for Sam to buy three servings, because the marginal utility (MU) of the fourth serving is $3.60. If price falls below $3.60, it then pays to purchase the fourth serving because its MU exceeds its price. The marginal utility of the next (fifth) serving is only $1.80, and so if the price were to remain above $1.80, it would not pay Sam to buy that fifth serving, just as prescribed in the optimal purchase rule.

Note the critical role of the law of diminishing marginal utility. If P falls, a consumer who wishes to maximize total utility must arrange things so that MU falls. According to the law of diminishing marginal utility, the only way to do this is to increase the quantity purchased.

While this explanation is a bit abstract, it can easily be rephrased in practical terms. We have noted that the various uses to which an individual puts a commodity have different priorities. For Sam, buying a portion of Maknoogats for his girlfriend has a higher priority than using the chicken to feed his roommate. If the price of Maknoogats is high, Sam will buy only enough for the high-priority uses—those that offer high marginal utilities. When price declines, however, it pays to purchase more of the good—enough for some low-priority uses. This is the essence of the analysis. The same assumption about consumer psychology underlies both the law of diminishing marginal utility and the negative slope of the demand curve. They are really two different ways of describing the assumed attitudes of consumers.

CONSUMER CHOICE AS A TRADE-OFF: OPPORTUNITY COST

We have expressed the optimal purchase rule as a decision about how much of *one* commodity to buy. However, we have already observed that the scarcity of income lurking in the background turns every such decision into a trade-off. Given the consumer's limited income, a decision to buy a new car may mean giving up some travel or postponing the purchase of some furniture. A purchase of a dozen videotapes may mean fewer trips to the movies. The money that the consumer gives up when she makes a purchase—her expenditure on that purchase—is only a measure of the true underlying cost to her. The real cost is the *opportunity cost* of the purchase—the commodities that she must give up as a result of the purchase decision.

Any decision to buy implies some such a trade-off because scarcity constrains all economic decisions. Even a multimillionaire faces very real trade-offs, even if they do not inspire much pity. Someone who decides to buy a factory may not be able to afford to invest in an office building.

This last example has one other important implication. The trade-off from a consumer's purchase decision does not always involve giving up another consumer good.

Consider, for example, the choice between consumption and saving. A decision to cut down on consumption now and put the money into the bank means that the consumer will be wealthier in the future because of the interest earned. This, in turn, will allow the person to afford more goods at some future date. But the opportunity cost of that enhanced future consumption will be the consumption given up today.

Thus, every purchase decision entails a trade-off. If the consumer is buying two goods, X and Y, optimality requires that the purchase of an *additional* dol-

lar's worth of X contribute just as much utility as a dollar's worth of Y. This is another way of saying that the opportunity cost incurred when a consumer spends an additional dollar on X is the utility of the amount of Y that the person would have gained by spending that dollar on Y instead.

CONSUMER'S SURPLUS: THE NET GAIN FROM A PURCHASE

CONSUMER'S SURPLUS is the difference between the value to the consumer of the quantity of Commodity X purchased and the amount that the market requires the consumer to pay for that quantity of X.

Our discussion of the optimal purchase rule, MU = P, assumed that the consumer was trying to maximize the money value of the total utility from the purchase *minus* the amount spent to make that purchase. Economists give the name **consumer's surplus** to that difference, that is, to the net gain in total utility that a purchase brings to a buyer. The consumer is trying to make the purchase decisions that maximize:

Consumer's surplus = Total utility (in money terms) − Total expenditure

Thus, just as economists assume that a business firm maximizes total profit (equal to total revenue minus total cost), they assume that the consumer maximizes consumer's surplus, that is, the difference between the total utility of the purchased commodity and the amount that the consumer spends on it.

The concept of *consumer's surplus* seems to suggest that the consumer gains some sort of free bonus, or *surplus*, for every purchase. In many cases, this idea seems absurd. How can it be true, particularly for goods whose prices seem to be outrageous?

We hinted at the answer in Chapter 3, where we observed that both parties must gain from a voluntary exchange or else one of them will refuse to participate. The same must be true when a consumer makes a *voluntary* purchase from a supermarket or an appliance store. If the consumer did not expect a net gain from the transaction, he or she would simply not bother to buy. Even if the seller were to "overcharge" by some standard, that would merely reduce the size of the consumer's net gain, not eliminate it entirely. If the seller is so greedy as to charge a price that wipes out the net gain altogether, the punishment will fit the crime: The consumer will refuse to buy, and the greedy seller's would-be gains will never materialize. The basic principle states that every purchase that is not marginal—that is, every purchase except those about which the consumer is indifferent—must yield *some* consumer's surplus.

But how large is that surplus? At least in theory, it can be measured with the aid of a table or graph of marginal utilities (Table 5-1 and Figure 5-1). Suppose that, as in our earlier example, the price of Maknoogats is \$3.60 per serving and Sam purchases four servings. Table 5-3 reproduces the marginal utility numbers from Table 5-1. It shows that the first serving is worth \$6.00 to Sam, so at the \$3.60 price, he reaps a net gain (surplus) of \$6.00 minus \$3.60, or \$2.40, by buying that serving. The second serving also brings Sam some surplus, but less than the first serving does, because the marginal utility diminishes. Specifically, the second portion provides a surplus of \$5.60 minus \$3.60, or \$2.00. Reasoning in the same way, the third serving gives Sam a surplus of \$4.40 minus \$3.60, or \$0.80. It is only the fourth serving—the last one that Sam purchases—that offers no surplus because, by the optimal purchase rule, the marginal utility of the last unit is equal to its price.

We can now easily determine the total consumer's surplus that Sam obtains by buying four units of Maknoogats. It is simply the sum of the surpluses

TABLE 5-3 CALCULATING MARGINAL NET UTILITY (CONSUMER'S SURPLUS) FROM SAM'S MAKNOOGAT PURCHASES

Quantity	Marginal Utility	Price	Marginal Net Utility (Surplus)
0			
1	$6.00	$3.60	$2.40
2	5.60	3.60	2.00
3	4.40	3.60	0.80
	3.60	3.60	0.00
Total			$5.20

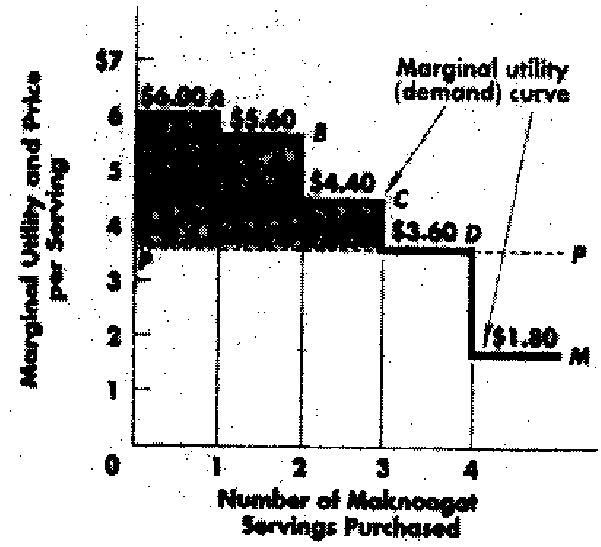
received from each serving. Table 5-3 shows that this total consumer's surplus is:

$$\$2.40 + \$2.00 + \$0.80 + \$0 = \$5.20$$

This way of looking at the optimal purchase rule shows why a buyer must always gain some consumer's surplus if she buys more than one unit of a good. Note that the price of each unit remains the same, but the marginal utility diminishes as more units are purchased. The last unit bought yields no consumer's surplus because MU = *P*. But all prior units must have had marginal utilities above the MU of the last unit because of diminishing marginal utility. Since each has the same cost, each must have yielded some positive consumer's surplus.

We can be more precise about the calculation of the consumer's surplus with the help of a graph (Figure 5-2) showing marginal utility as a set of bars. The bars labeled *A, B, C,* and *D* come from the corresponding points on Sam's marginal utility curve (demand curve) in Figure 5-1 (page 99). In Figure 5-2, the bar whose upper, right-hand corner is labeled *A* represents the $6.00 marginal utility Sam derives from the first serving of Maknoogats; the same interpretation applies to the bars *B, C,* and *D*. The horizontal line *PP* connotes the (fixed) $3.60 price.

FIGURE 5-2 GRAPHIC CALCULATION OF CONSUMER'S SURPLUS



Sam's consumer's surplus from each serving of Maknoogats he purchases equals the marginal utility of that serving minus the price he pays for it. By representing consumer's surplus graphically, one can determine just how much Sam obtains from his entire purchase by measuring the area between the marginal utility curve and the horizontal line representing the price of Maknoogats.

Clearly, the first serving that Sam purchases yields a consumer's surplus of $2.40, indicated by the shaded part of bar *A*. The height of that part of the bar is equal to the $6.00 marginal utility minus the $3.60 price. In the same way, the other two shaded areas represent the surpluses offered by the second and third servings of Maknoogats. The fourth serving has no shaded area because the height representing marginal utility is equal to the height representing price. Sum the shaded areas in the graph to obtain, once again, Sam's total consumer's surplus ($2.40 + $2.00 + $0.80 = $5.20) from his four-serving purchase.

The consumer's surplus derived from buying a certain number of units of a good is obtained graphically by drawing the person's demand curve as a set of bars whose heights represent the marginal utilities of the corresponding quantities of the good, and then drawing a horizontal line whose height is the price of the good. The sum of the heights of the bars above the horizontal line, that is, the area of the demand (marginal utility) bars above that horizontal line, measures the *total* consumer's surplus that the purchase yields.

RESOLVING THE DIAMOND-WATER PARADOX

We can now use marginal utility analysis to solve Adam Smith's paradox (which he was never able to explain) that diamonds are very expensive while water is generally very cheap, even though water seems to offer far more utility. The resolution of the diamond-water paradox is based on the distinction between marginal and total utility.

The *total* utility of water—its role as a necessity of life—is indeed much higher than that of diamonds. But price, as we have seen, is not related directly to *total* utility. Rather, the optimal purchase rule tells us that price tends to equal *marginal* utility. We have every reason to expect the marginal utility of water to be very low while the marginal utility of a diamond is very high.

Water is comparatively cheap to provide, so its price is generally quite low. Consumers thus use correspondingly large quantities of water. The principle of diminishing marginal utility, therefore, pushes down the marginal utility of water for a typical household to a low level. As the consumer's surplus diagram (Figure 5-2) suggests, this also means that its *total* utility is likely to be high.

On the other hand, high-quality diamonds are scarce (partly because a monopoly keeps them so). As a result, the quantity of diamonds consumed is not large enough to drive down the MU of diamonds very far, so buyers of such luxuries must pay high prices for them. As a commodity becomes more scarce, its *marginal* utility and its market price rise higher, regardless of the size of its *total* utility. Also, as we have seen, because so little of the commodity is consumed, its total utility is likely to be comparatively low, despite its large marginal utility.

Thus, like many paradoxes, the diamond-water puzzle has a straightforward explanation. In this case, all one has to remember is that:

Scarcity raises price and *marginal* utility, but it generally reduces *total* utility.

PRICE, INCOME, AND QUANTITY DEMANDED

Our application of marginal analysis has enabled us to examine the relation between the *price* of a commodity and the quantity that will be purchased. Let us next consider briefly how quantity demanded responds to a change in *income*.

THE CONSEQUENCES OF A CHANGE IN INCOME

As a concrete example, consider what happens to the number of ball-point pens a consumer will buy when her real income rises. It may seem almost certain that she will buy more such pens than before, but that is not necessarily so. A rise in real income can either increase or decrease the quantity purchased.

Why might an increase in income lead a consumer to buy fewer ball-point pens? People buy some goods and services only because they cannot afford anything better. They eat chicken three days a week and lobster twice a year, but they would rather have it the other way around. They may use ball-point pens instead of finely crafted fountain pens, or purchase clothing secondhand instead of new. If their real income rises, they may then buy more lobster and less chicken, more fountain pens and fewer ball points, and perhaps more steaks and fewer hamburgers. Thus, a rise in real income will reduce the quantities of chicken, cheap pens, and possibly even hamburgers demanded. Economists have given the rather descriptive name **inferior goods** to the class of commodities for which quantity demanded falls when income rises.

An **INFERIOR GOOD** is a commodity whose quantity demanded falls when the purchaser's real income rises, all other things remaining equal.

The upshot of this discussion is that economists cannot draw definite conclusions about the effects of a rise in consumer incomes on quantity demanded. But for most commodities, if incomes rise and prices do not change, quantity demanded will increase. (Such an item is often called a *normal good.*)

THE TWO EFFECTS OF A CHANGE IN PRICE[4]

When the price of some good, say, heating oil, falls, it has two consequences. First, the drop makes fuel oil cheaper relative to electricity, gas, or coal. Economists say, then, that the *relative price* of fuel oil has fallen. Second, this price decrease leaves homeowners with more money to spend on movie tickets, soft drinks, or clothing. In other words, the decrease in the price of fuel oil *increases the consumer's real income,* that is, the power to purchase other goods.

The **INCOME EFFECT** is a *portion* of the change in quantity of a good demanded when its price changes. A rise in price cuts the consumer's purchasing power (real income), which leads to a change in the quantity demanded of that commodity. That change is the income effect.

While a fall in the price of a commodity always produces these two effects simultaneously, our analysis will be easier if we separate the effects from one another and study them one at a time.

1. *The income effect of a change in price.* As we have just noted, a fall in the price of a commodity leads to a rise in the consumer's *real* income—the amount that her income will purchase. The consequent effect on quantity demanded is called the **income effect** of the price fall. The income effect caused by a fall in a commodity's price is much the same as the effect of a rise in the consumer's income: She will buy more of any commodity that is not an inferior good. The process that produces the income effect thus has three stages: (1) the price of the good falls, causing (2) an increase in the consumer's real income, which leads to (3) a change in quantity demanded. Of course, a price increase will produce the same effect in reverse. The consumer's real income will decline, leading to the opposite change in quantity demanded.
2. *The substitution effect of a change in price.* A change in the price of a commodity produces another effect on quantity demanded that is rather different from the income effect. This **substitution effect** is the effect on quantity demanded attributable to the fact that the new price is now higher or lower

The **SUBSTITUTION EFFECT** is the change in quantity demanded of a good resulting from a change in its *relative* price, exclusive of whatever change in quantity demanded may be attributable to the associated change in real income.

[4]This section contains rather more difficult material, which, in shorter courses, may be omitted without loss of continuity.

than before *relative to the prices of other goods*. The substitution effect of a price change is the portion of the change in quantity demanded that can be attributed *exclusively* to the change in relative prices rather than to the associated change in real income.

There is nothing mysterious or surprising about the effect of a change in relative prices when the consumer's real income remains unchanged. Whenever the consumer can switch between two commodities, she can be expected to buy more of the good whose relative price has fallen and less of the good whose relative price has risen. For example, a few years ago, AT&T sharply reduced the price of evening long-distance telephone calls relative to the price of daytime calls. The big decrease in the relative price of evening calls brought about a large increase in calling during the evening hours and a decrease in daytime calling, just as the telephone company had hoped. Similarly, many utility companies flatten the peaks and valleys in the demand for electricity by offering customers cheaper rates during off-peak hours.

When the price of any Commodity X rises relative to the price of some other Commodity Y, a consumer whose real income has remained unchanged can be expected to buy less X and more Y than before. Thus, *considering the substitution effect alone*, a decline in price always increases quantity demanded and a rise in price always reduces quantity demanded.

Beginning economics students may suspect that the income effect and the substitution effect were invented just to torture them. In fact, the concepts are really quite useful. Consider this example. Suppose that the price of hamburgers declines while the price of cheese remains unchanged. The *substitution effect* clearly induces the consumer to buy more hamburgers in place of grilled-cheese sandwiches, because hamburgers are now comparatively cheaper. What about the *income effect?* Unless a hamburger is an inferior good, it leads to the same decision. The fall in price makes consumers richer, which induces them to increase their purchases of all but inferior goods. This example illustrates two general points:

If a good is not inferior, it must have a downward-sloping demand curve, since the income and substitution effects reinforce each other. However, an inferior good may violate this pattern of demand behavior because the income effect of a decline in price leads consumers to buy less.

Do all inferior goods, then, have upward-sloping demand curves? Certainly not! Remember the substitution effect; it always favors a downward-sloping demand curve. Thus, an inferior good creates a kind of tug-of-war. If the *income effect* predominates, the demand curve will slope upward; if the *substitution effect* prevails, the demand curve will slope downward.

Economists have concluded that the substitution effect generally wins out; so while the market includes many examples of inferior goods, it includes few examples of upward-sloping demand curves. When might the income effect prevail over the substitution effect? Certainly not when the good in question (say, plastic wallets) accounts for a very small fraction of the consumer's budget, for then a fall in price makes the consumer only slightly richer and therefore creates a very small income effect. But the demand curve can slope upward if an inferior good constitutes a substantial portion of the consumer's budget.

We conclude this discussion of income and substitution effects with a warning against a frequently made error. Many students mistakenly close their books thinking that price changes cause substitution effects while income changes cause income effects. This is incorrect. As the example of hamburgers made clear:

Any change in price sets in motion *both* a substitution effect *and* an income effect, each of which affects quantity demanded.

FROM INDIVIDUAL DEMAND CURVES TO MARKET DEMAND CURVES

So far in this chapter we have studied how *individual demand curves* are derived from the logic of consumer choice. But to understand how the market system works, we must derive the relationship between price and quantity demanded *in the market as a whole*—the **market demand curve.**

A **MARKET DEMAND CURVE** shows how the total quantity demanded of some product during a specified period of time changes as the price of that product changes, holding other things constant.

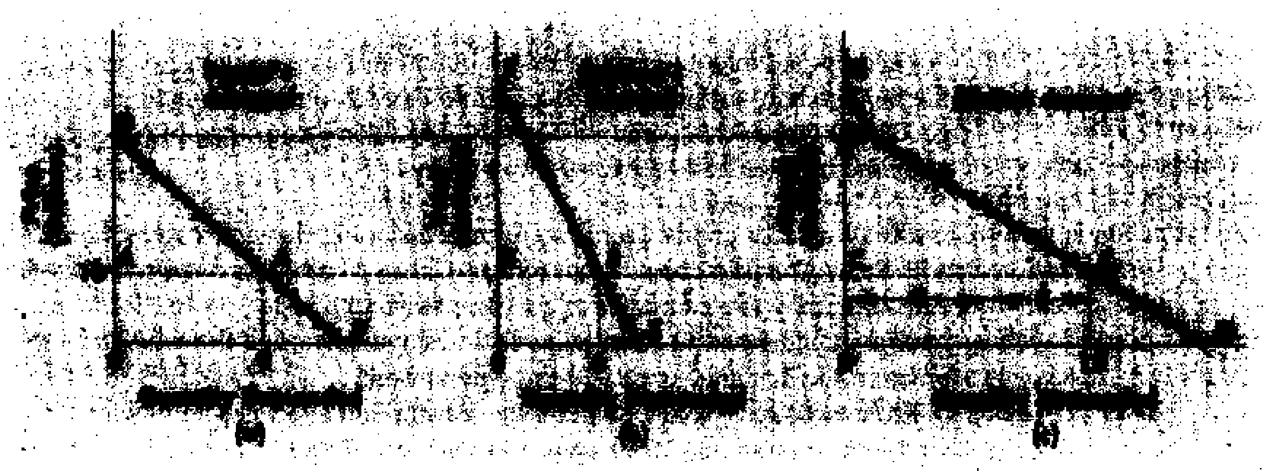
If each individual pays no attention to other people's purchase decisions when making his own, it is easy to derive the market demand curve from consumers' individual demand curves: Simply *add* the negatively sloping individual demand curves *horizontally* as shown in Figure 5-3. The figure gives the individual demand curves *DD* and *ZZ* for two people, Daniel and Sabrina, and the total (market) demand curve, *MM*.

Specifically, we have derived this market demand curve in the following straightforward way.

Step 1: Pick any relevant price, say, $10.

Step 2: At that price, determine Daniel's quantity demanded (9 units) from his demand curve in Panel (a) and Sabrina's quantity demanded (6 units) from her demand curve in Panel (b). Note that these quantities are indicated by the line segment labeled *AA* for Daniel and that labeled *BB* for Sabrina.

FIGURE 5-3 THE RELATIONSHIP BETWEEN TOTAL MARKET DEMAND AND THE DEMAND OF INDIVIDUAL CONSUMERS WITHIN THAT MARKET



Daniel and Sabrina are the customers for a product. At a price of $10, Daniel demands 9 units, as line *AA* in Panel (a) shows; Sabrina demands 6 units, as shown by line *BB* in Panel (b), so total quantity demanded by the market at that price is 9 plus 6, or 15 units, indicated by line *CC* in Panel (c). In other words, we obtain the market demand curve by adding horizontally all points on each consumer's demand curve at each given price. Thus, at a $10 price, market demand equals the length of *CC* on the market demand curve, which is equal to *AA* plus *BB* on the individual demand curves. (The sharp angle at point *K* on the market curve occurs because that point corresponds to the price at which Daniel, whose demand pattern is different from Sabrina's, first enters the market. At any higher price, only Sabrina is willing to buy anything.)

Step 3: Add Sabrina's and Daniel's quantities demanded at the \$10 price (segment AA + segment $BB = 9 + 6 = 15$) to yield the total quantity demanded by the market at that price. This gives segment CC, with total quantity demanded equal to 15 units, in Panel (c).

Now repeat the process for all alternative prices to obtain other points on the market demand curve until the shape of the entire curve MM is indicated. That is all there is to the adding-up process. (Question: What would happen to the market demand curve if, say, another consumer entered the market?)

THE "LAW" OF DEMAND

The **"LAW" OF DEMAND** states that a lower price generally increases the amount of a commodity that people in a market are willing to buy. Therefore, for most goods, demand curves have negative slopes.

As for an individual's demand curve, we expect the total quantity demanded by the market to move in the opposite direction from price. Economists call this relationship the **"law" of demand.**

Notice that we have put the word *law* in quotation marks. By now you will have observed that economic laws are not always obeyed, and we shall see in a moment that the "law" of demand is not without its exceptions. But first let us see why the "law" usually holds.

Earlier in this chapter, we explained that individual demand curves usually slope downward because of the "law" of diminishing marginal utility. If individual demand curves slope downward, then the preceding discussion of the adding-up process implies that the market demand curve must also slope downward. This is just common sense; if every consumer in the market buys less chicken when the price of chicken rises, then the total quantity demanded in the market must surely fall.

But market demand curves may slope downward even when individual demand curves do not, because not all consumers are alike. For example, if a bookstore reduces the price of a popular novel, it may draw many new customers, but few of the customers who already own a copy will be induced to buy a second one. Similarly, people differ in their fondness for chicken. True devotees may maintain their purchases of chicken even at exorbitant prices, while others would not eat chicken even if it were offered free of charge. As the price of chicken rises, the less enthusiastic chicken eaters drop out of the market entirely, leaving the expensive meat to the more devoted consumers. Thus, the quantity demanded declines as price rises simply because higher prices induce more people to kick the chicken habit. Indeed, for many commodities, it is the appearance of new customers in the *market* when prices are lower, rather than the negative slope of *individual* demand curves, that accounts for the law of demand.

This is also illustrated in Figure 5-3, where only Sabrina will buy the product at a price higher than D. However, at a price below D, Daniel is also induced to make some purchases. Hence, below point K the market demand curve lies further to the right than it would have if Daniel had not been induced to enter the market. Put another way, a rise in price from a level below D to a level above D would cut quantity demanded for two reasons: first, because Sabrina's demand curve has a negative slope and, second, because it would drive Daniel out of the market.

We conclude, therefore, that the law of demand stands on fairly solid ground. If individual demand curves are downward sloping, then the market demand curve surely will be, too. Further, the market demand curve may slope downward, even when individual demand curves do not.

Nevertheless, exceptions to the law of demand have been noted. One common exception occurs when people judge quality on the basis of price—they perceive a more expensive commodity as offering better quality. For example, many people buy name-brand aspirin, even if right next to it on the drugstore shelf they see an unbranded, generic aspirin with an identical chemical formula and selling at half the price. The consumers who do buy the name-brand aspirin may well use comparative price to judge the relative qualities of different brands. They may prefer Brand *X* to Brand *Y* because *X* is slightly more expensive. If Brand *X* were to reduce its price below that of *Y*, consumers might assume that it was no longer superior and actually reduce their purchases.

Another possible cause of an upward-sloping demand curve is snob appeal. If part of the reason for purchasing a Rolls Royce is to advertise one's wealth, a decrease in the car's price may actually reduce sales, even if the quality of the car remains unchanged. Other types of exceptions have also been noted by economists. But, for most commodities, it seems quite reasonable to assume that demand curves have negative slopes, an assumption that is supported by the data.

SUMMARY

1. Economists distinguish between total and marginal utility. **Total utility**, or the benefit a consumer derives from a purchase, is measured by the maximum amount of money he or she would give up in order to have the good. Rational consumers seek to maximize (net) total utility, or, **consumer's surplus:** the total utility derived from a commodity minus the value of the money spent in buying it.

2. **Marginal utility** is the maximum amount of money that a consumer is willing to pay for an additional unit of a particular commodity. Marginal utility is useful in calculating the set of purchases that maximizes net total utility.

3. **The "law" of diminishing marginal utility** is a psychological hypothesis stating that, as a consumer acquires more and more of a commodity, the marginal utility of additional units of the commodity decreases.

4. To maximize the total utility obtained by spending money on some Commodity *X*, given the fact that other goods can be purchased only with the money that remains after buying *X*, the consumer must purchase a quantity of *X* such that the price equals the commodity's marginal utility (in money terms).

5. If the consumer acts to maximize utility, and if his marginal utility of some good declines when he purchases larger quantities, then his demand curve for the good will have a negative slope. A reduction in price will induce him to purchase more units, leading to a lower marginal utility.

6. Abundant goods tend to have low prices and low marginal utilities regardless of whether their total utilities are high or low. That is why water can have a lower price than diamonds despite its higher total utility.

7. An **inferior good**, such as secondhand clothing, is a commodity that consumers buy less of when they get richer, all other things held equal.

8. Consumers usually earn a surplus when they purchase a commodity voluntarily. This means that the quantity of the good that they buy is worth more to them than the money they give up in exchange. Consumer's surplus is normally positive.

9. A rise in the price of a commodity has two effects on quantity demanded: (a) a **substitution effect**, which makes the good less attractive because it has become more expensive than it was previously, and (b) an **income effect**, which affects the quantity of the good the consumer buys because higher prices cut her purchasing power.

10. Any increase in the price of a good always has a *negative* substitution effect; that is, considering only the substitution effect, a rise in price must reduce the quantity demanded.

11. The income effect of a rise in price may, however, push quantity demanded up or down. For normal

goods, the income effect of a higher price (which makes consumers poorer) reduces the quantity demanded; for inferior goods, the income effect of higher prices actually increases the quantity demanded.

12. The demand curve for an entire market is obtained by taking a horizontal sum of the demand curves of all of the individuals who buy or consider buying in that market. This sum is obtained by adding up, for each price, the quantity of the commodity in question that every such consumer is willing to purchase at that price.

KEY TERMS

Marginal analysis
Total utility
Marginal utility
The "law" of diminishing marginal utility
Optimal purchase rule
Scarcity and marginal utility
Consumer's surplus
Inferior good
Income effect
Substitution effect
Market demand curve
The "law" of demand

QUESTIONS FOR REVIEW

1. Describe some of the different things you do with water. Which would you give up if the price of water were to rise a little? If it were to rise by a fairly large amount? If it were to rise by a very large amount?
2. Which gives you greater *total* utility, 12 gallons of water per day or 20 gallons per day? Why?
3. At which level do you get greater *marginal* utility, at 12 gallons per day or 20 gallons per day? Why?
4. Suppose that you wanted to measure the marginal utility of a commodity to a consumer by directly determining the consumer's psychological attitude or strength of feeling for the commodity rather than by seeing how much money the consumer would give up for the commodity. Why would you find it difficult to get such a psychological measurement?
5. Some people who do not understand the optimal purchase rule argue that if a consumer buys so much of a good that its price equals its marginal utility, she could not possibly be behaving optimally. Rather, they say, she would be better off quitting while she was ahead, that is, buying a quantity such that marginal utility is much greater than price. What is wrong with this argument? (*Hint:* What opportunity would the consumer then miss? Is it maximization of marginal or total utility that serves the consumer's interests?)
6. What inferior goods do you purchase? Why do you buy them? Do you think you will continue to buy them when your income is higher?
7. Which of the following items are likely to be normal goods for a typical consumer? Which are likely to be inferior goods?
   a. Expensive perfume
   b. Paper plates
   c. Secondhand clothing
   d. Overseas trips
8. Suppose that electricity and paper clips each rise in price by 40 percent. Which will have the larger income effect on the purchases of a typical consumer? Why?
9. Around 1850, Sir Robert Giffen observed that Irish peasants actually consumed more potatoes as the price of potatoes increased. Use the concepts of the income and substitution effects to explain this phenomenon.
10. Suppose that strawberries sell for $2 per basket. Jim is considering whether to buy 0, 1, 2, 3, or 4 baskets. Construct a table of total and marginal utilities for the different quantities of strawberries and calculate how many baskets Jim would buy.
11. Draw a graph showing the consumer's surplus Jim would get from his strawberry purchase and check your answer with the help of your marginal utility table.
12. Consider a market with two consumers, Jasmine and Jim. Draw a demand curve for each of the two consumers and use those curves to construct the demand curve for the entire market.

APPENDIX INDIFFERENCE CURVE ANALYSIS

The analysis of consumer demand presented in this chapter, while correct as far as it goes, has one shortcoming: By treating the consumer's decision about the purchase of each commodity as an isolated event, it conceals the necessity of choice imposed on the consumer by his limited budget. It does not indicate explicitly the hard choice behind every purchase decision—the sacrifice of some goods to obtain others.

The idea is included implicitly, of course, because the purchase of a commodity involves a trade-off between that good and money. If you spend more money on rent, you have less to spend on entertainment. If you buy more clothing, you have less money for food. But to represent the consumer's *choice* problem explicitly, economists have invented two geometric devices, the *budget line* and the *indifference curve*, which are described in this appendix.

GEOMETRY OF AVAILABLE CHOICES: THE BUDGET LINE

Suppose, for simplicity, that only two commodities were produced in the world, cheese and rubber bands. The decision problem of any household then would be to determine the allocation of its income between these two goods. Clearly, the more it spends on one, the less it can have of the other. But just what is the trade-off? A numerical example will answer this question and also introduce the graphical device that economists use to portray the trade-off.

Suppose that cheese costs $2 per pound, boxes of rubber bands sell at $3 each, and a consumer has $12 at her disposal. She obviously has a variety of choices, as displayed in Table 5-4. For example, if she buys no rubber bands, she can go home with 6 pounds of cheese, and so on. Each of the combinations of cheese and rubber bands that the consumer can afford can be shown in a diagram in which the axes measure the quantities purchased of each commodity. In Figure 5-4, pounds of cheese are measured along the vertical axis, the number of boxes of rubber bands is measured along the horizontal axis, and a labeled point represents each of the combinations enumerated in Table 5-4. For example, point *A* corresponds to spending everything on cheese, point *E* corresponds to spending everything on rubber bands, and point *C* corresponds to buying two boxes of rubber bands and 3 pounds of cheese, which together use up the $12 available.

If a straight line connects points *A* through *E*, the orange line in the diagram, it traces all possible ways to divide the $12 between the two goods. For example, at point *D*, if the consumer buys three boxes of rubber bands, she will have only enough money left to purchase 1-1/2 pounds of cheese. This is readily seen to be correct from Table 5-4. Line *AE* is therefore called the *budget line.*

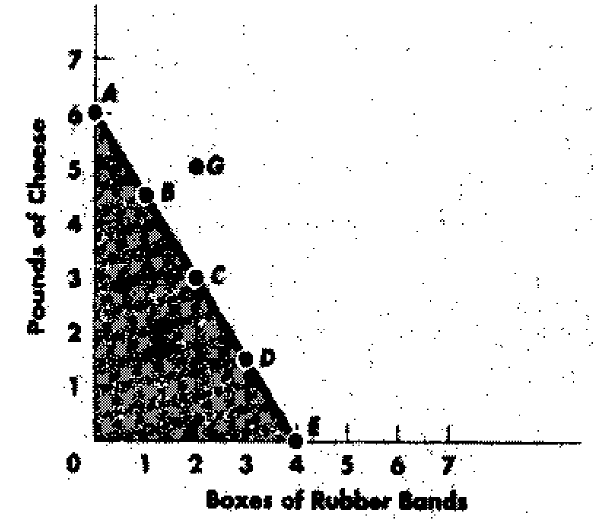
The **budget line** for a household represents graphically all possible combinations of two commodities that it can purchase, given the prices of the commodities and some fixed amount of money at its disposal.

PROPERTIES OF THE BUDGET LINE Let us now use *r* to represent the number of boxes of rubber bands

TABLE 5-4 ALTERNATIVE PURCHASE COMBINATIONS FOR A $12 BUDGET

Boxes of Rubber Bands (at $3 each)	Expenditure on Rubber Bands (in dollars)	Remaining Funds (in dollars)	Pounds of Cheese (at $2 each)	Label in Figure 5-4
0	0	12	6	*A*
1	3	9	4½	*B*
2	6	6	3	*C*
3	9	3	1½	*D*
4	12	0	0	*E*

FIGURE 5-4 A BUDGET LINE



This budget line shows the different combinations of cheese and rubber bands that the consumer can buy with \$12 if cheese costs \$2 per pound and a box of rubber bands costs \$3. At point *A*, the consumer buys 6 pounds of cheese and has nothing left over for rubber bands. At point *E*, she spends the entire budget on rubber bands. At intermediate points on the budget line (such as *C*), the consumer buys some of both goods (two boxes of rubber bands and 3 pounds of cheese).

purchased by the consumer and c to indicate the amount of cheese that she acquires. Thus, at \$2 per pound, she spends on cheese a total of \$2 times the number of pounds of cheese bought, or $\$2c$. Similarly, she spends $\$3r$ on rubber bands, making a total of $\$2c$ plus $\$3r$, which must equal \$12, if she spends the entire \$12 on the two commodities. Thus, $2c + 3r = 12$ is the equation of the budget line. It is also the equation of the straight line drawn in the diagram.[5]

Note also that the budget line represents the *maximum* amounts of the commodities that the consumer can afford. Thus, for any given purchase of rubber bands, it indicates the greatest amount of cheese that her money can buy. If the consumer wants to be thrifty, she can choose to end up at a point below the budget line, such as *K*. Clearly, then, the choices she has available include not only those points on the budget line, *AE*, but also any point in the shaded triangle formed by that line and the two axes. By contrast, points above the budget line, such as *G*, are not available to the consumer given her limited budget. A bundle of 5 pounds of cheese and two boxes of rubber bands would cost \$16, which is more than she has to spend.

The position of the budget line is determined by two types of data: the prices of the commodities purchased and the income at the buyer's disposal. We can complete our discussion of the graphics of the budget line by examining briefly how a change in either of these magnitudes affects the location of that line.

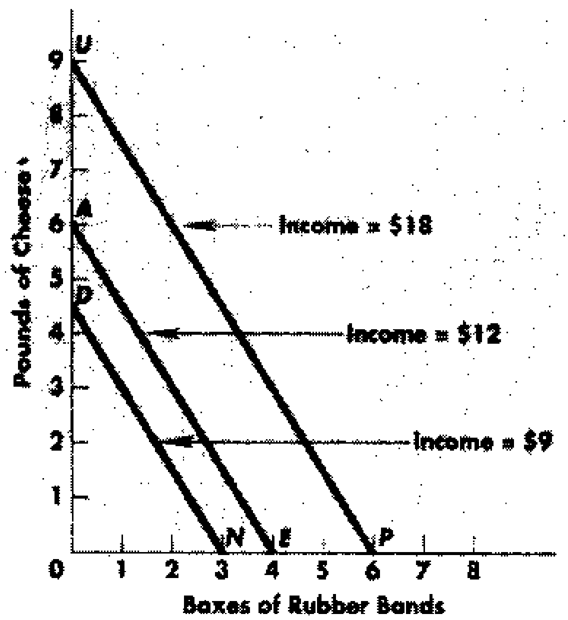
Obviously, any increase in the income of the household increases the range of options available to it. Specifically, *increases in income produce parallel shifts in the budget line*, as shown in Figure 5-5. The reason is simply that an increase in available income of, say, 50 percent, if spent entirely on these two goods, would permit the family to purchase exactly 50 percent more of *either* commodity. Point *A* in Figure 5-4 would shift upward by 50 percent of its distance from the origin, while point *E* would move to the right by 50 percent.[6] Figure 5-5 shows

[5]The reader may have noticed one problem that arises in this formulation. If every point on the budget line, *AE*, is a possible way for the consumer to spend her money, she must be able to buy fractional boxes of rubber bands. Perhaps the purchase of 1 and 1/2 boxes can be interpreted to include a down payment of \$1.50 on a box of rubber bands to be purchased on her next shopping trip! Throughout this book it is convenient to assume that commodities are available in fractional quantities when drawing diagrams. This makes the graphs clearer and does not really affect the analysis.

[6]An algebraic proof is simple. Let M (which is initially \$12) be the amount of money available to the household. The equation of the budget line can be solved for c, obtaining $c = -(3/2)r + M/2$. This is the equation of a straight line with a slope of $-3/2$ and a vertical intercept of $M/2$. A change in M, the quantity of money available, will not change the *slope* of the budget line; it will lead only to parallel shifts in that line.

FIGURE 5-5 THE EFFECT OF INCOME CHANGES ON THE BUDGET LINE

A change in the amount of money in the consumer's budget causes a parallel shift in the budget line. A rise in the budget from $12 to $18 raises the budget line from *AE* to *UP*. A fall from $12 to $9 lowers the budget line from *AE* to *DN*.



three such budget lines corresponding to incomes of $9, $12, and $18, respectively.

Finally, we can ask what happens to the budget line when the price of some commodity changes. In Figure 5-6, when the price of the rubber bands *decreases*, the budget line moves outward, but the move is no longer parallel because the point on the cheese axis remains fixed. Once again, the reason is fairly straightforward. A 50 percent reduction in the price of rubber bands permits the consumer to buy twice as many boxes of rubber bands with her $12 as before: point *E* moves rightward to point *H*, where the buyer can obtain eight boxes of rubber bands. However, since the price of cheese has not changed from point *A*, the amount of cheese that can be bought for $12 is unaffected. This gives the general result that *a reduction in the price of one of the two commodities swings the budget line outward along the axis representing the quantity of that item while leaving the location of the other end of the line unchanged.*

WHAT THE CONSUMER PREFERS: THE INDIFFERENCE CURVE

The budget line indicates what choices are *available* to the consumer, given the size of her income and the commodity prices fixed by the market. Next, we must examine the consumer's *preferences* in order to determine which of these available possibilities she will want to choose.

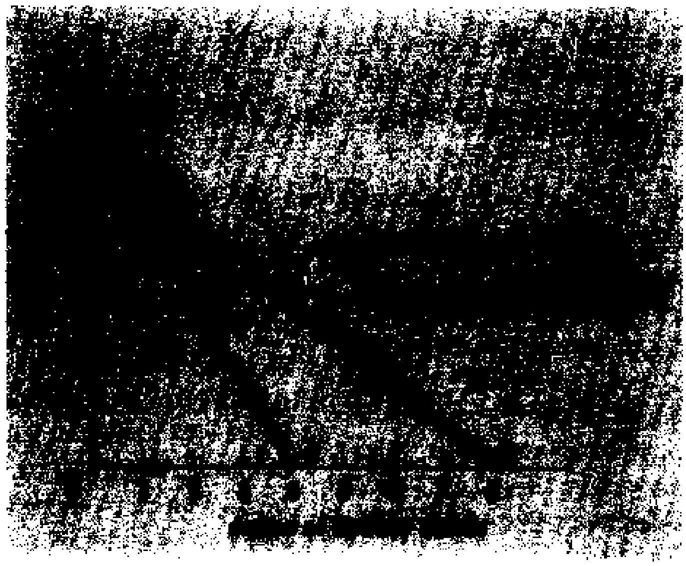
After much investigation, economists have determined what they believe to be the minimum amount of information they need about a purchaser in order to analyze her choices. This information consists of the consumer's *ranking* of the alternative bundles of commodities that are available. Suppose, for instance, that the consumer can choose between two bundles of goods, Bundle *W*, which contains three boxes of rubber bands and 1 pound of cheese, and Bundle *T*, which contains two boxes of rubber bands and 3 pounds of cheese. The economist wants to know for this purpose only whether the consumer prefers *W* to *T* or *T* to *W*, or whether she is *indifferent* about which one she gets. Note that the analysis requires no information about *degree* of preference—whether the consumer is wildly more enthusiastic about one of the bundles or just prefers it slightly.

Graphically, the preference information is provided by a group of curves called *indifference curves* (Figure 5-7).

An **indifference curve** is a line connecting all combinations of the commodities that are equally desirable to the consumer.

FIGURE 5-6 THE EFFECT OF PRICE CHANGES ON THE BUDGET LINE

A fall in the price of rubber bands causes the end of the budget line on the rubber bands axis to swing away from the origin. A fall from $3.00 to $1.50 swings the price line from *AE* to orange line *AH*. This happens because at the higher price, $12 buys only four boxes of rubber bands, but at the lower price, it can buy eight boxes.

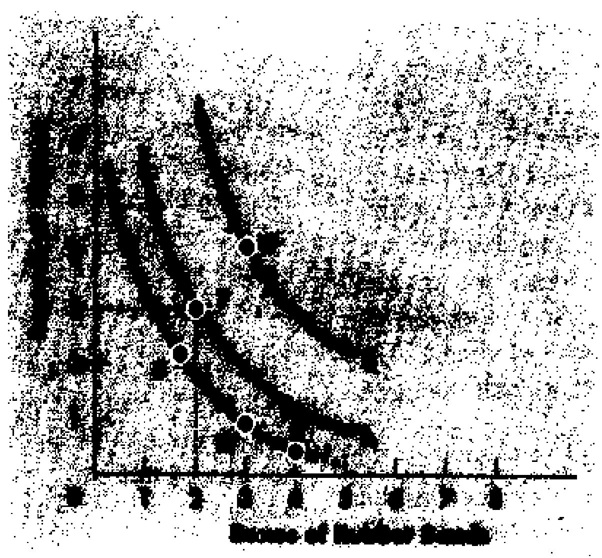


But before we examine these curves, let us see how such a curve is interpreted. A single point on an indifference curve says nothing about preferences. For example, point *R* on curve *I* simply represents the bundle of goods composed of four boxes of rubber bands and 1/2 pound of cheese. It does *not* suggest that the consumer is indifferent between 1/2 pound of cheese and four boxes of rubber bands. For the curve to indicate anything, one must consider at least two of its points, for example, points *S* and *W*. An indifference curve is defined to represent all such combinations that are equally desirable to the consumer.

PROPERTIES OF THE INDIFFERENCE CURVE No one knows yet which bundle, among all of the bundles she can afford, the consumer prefers; this analysis indicates only that a choice between certain bundles will lead to indifference. Before using indifference curves to analyze the consumer's choice, one

FIGURE 5-7 THREE INDIFFERENCE CURVES FOR CHEESE AND RUBBER BANDS

Any point in the diagram represents a combination of cheese and rubber bands. (For example, point *T* on indifference curve I_b represents two boxes of rubber bands and 3 pounds of cheese.) Any two points on the same indifference curve (for example, *S* and *W*, on indifference curve I_a) represent two combinations of the goods that the consumer likes equally well. If two points, such as *T* and *W*, lie on different indifference curves, the consumer prefers the one on the *higher* indifference curve.



must examine a few of its properties. Most important is the fact that:

As long as the consumer desires *more* of each of the goods in question, *every* point on a higher indifference curve (that is, a curve farther from the origin in the graph) will be preferred to *any* point on a lower indifference curve.

In other words, among indifference curves, higher is better. The reason is obvious. Given two indifference curves, say, I_b and I_c in Figure 5-7, the higher curve will contain points lying above and to the right of some points on the lower curve. Thus, point *U* on curve I_c lies above and to the right of point *T* on curve I_b. This means that the consumer gets more rubber bands *and* more cheese at *U* than at *T*. Assuming that she desires both commodities, the consumer must prefer *U* to *T*.

Since every point on curve I_c is, by definition, equal in desirability to point *U*, and the same relation holds for point *T* and all other points along curve I_b, the consumer will prefer *every* point on curve I_c to *any* point on curve I_b.

This at once implies a second property of indifference curves: they never intersect. This is so because if an indifference curve, say, I_b, is anywhere above another, say, I_a, then I_b must be above I_a everywhere, since every point on I_b is preferred to every point on I_a.

Another property that characterizes the indifference curve is its *negative slope*. Again, this holds only if the consumer wants more of both commodities. Consider two points, such as *S* and *R*, on the same indifference curve. If the consumer is indifferent between them, one point cannot represent more of *both* commodities than the other point. Since point *S* represents more cheese than point *R*, *R* must offer more rubber bands than *S*, or the consumer would not be indifferent about which she gets. This means that any movement toward the one with the larger number of rubber bands implies a decrease in the quantity of cheese. The curve will always slope downhill toward the right, a negative slope.

A final property of indifference curves is the nature of their curvature—the way they round toward the axes. They are drawn "bowed in"—they flatten out (their slopes decrease in absolute value) as they extend from left to right. To understand why this is so, we must first examine the economic interpretation of the slope of an indifference curve.

THE SLOPES OF INDIFFERENCE CURVES AND BUDGET LINES

In Figure 5-8, the average slope of the indifference curve between points *M* and *N* is represented by *RM/RN*. *RM* is the quantity of cheese that the consumer gives up in moving from *M* to *N*. Similarly, *RN* is the increased number of boxes of rubber bands acquired in this move. Since the consumer is indifferent between bundles *M* and *N*, the gain of *RN* rubber bands must just suffice to compensate her for the loss of *RM* pounds of cheese. Thus, the ratio *RM/RN* represents the terms on which the consumer is just willing—*according to her own preference*—to trade one good for the other. If *RM/RN* equals 2, the consumer is willing to give up (no more than) 2 pounds of cheese for one additional box of rubber bands.

The **slope of an indifference curve**, referred to as the **marginal rate of substitution** between the commodities, represents the maximum amount of one commodity that the consumer is willing to give up in exchange for one more unit of another commodity.

The slope of the budget line, *BB*, in Figure 5-8 is also a rate of exchange between cheese and rubber bands. But it no longer reflects the consumer's subjective willingness to trade. Rather, the slope represents the rate of exchange that *the market* offers to the consumer when she gives up money in exchange for cheese and rubber bands. Recall that the budget line represents all commodity combinations that a consumer can get by spending a fixed amount of money. The budget line is thus a curve of constant expenditure. At current prices, if the consumer reduces her purchase of cheese by amount *DE* in Figure 5-8, she will save just enough money to buy an additional amount, *EF*, of rubber bands, since at points *D* and *F* she is spending the same total number of dollars.

The **slope of a budget line** is the amount of one commodity that the market requires an individual to give up in order to obtain one additional unit of another commodity without any change in the amount of money spent.

The slopes of the two types of curves, then, are perfectly analogous in their meaning. The slope of the indifference curve indicates the terms on which the *consumer* is willing to trade one commodity for another, while the slope of the budget line reports

FIGURE 5-8 SLOPES OF A BUDGET LINE AND AN INDIFFERENCE CURVE

The slope of the budget line shows how many pounds of cheese, *ED*, the consumer can exchange in the market for *EF* boxes of rubber bands. The slope of the indifference curve shows how many pounds of cheese, *RM*, the consumer is just willing to exchange for *RN* boxes of rubber bands. When the consumer has more rubber bands and less cheese (point *m* rather than *M*), the slope of the indifference curve decreases, meaning that the consumer is only willing to give up *rm* pounds of cheese for *rn* boxes of rubber bands.



the *market* terms on which the consumer can trade one good for another.

It is useful to carry our interpretation of the slope of the budget line one step further. Common sense suggests that the market's rate of exchange between cheese and rubber bands should be related to their prices, *pc* and *pr*, and it is easy to show that this is so. Specifically, the slope of the budget line is equal to the ratio of the prices of the two commodities. The reason is straightforward. If the consumer gives up one box of rubber bands, she has *pr* more dollars to spend on cheese. But at a lower price of cheese, this money will enable her to buy a greater quantity of cheese—that is, her cheese purchasing power will be inversely related to its price. Since the price of cheese is *pc* per pound, the additional *pr* dollars permit her to buy *pr*/*pc* more pounds of cheese. Thus, the slope of the budget line is *pr*/*pc*.

Before returning to our main subject, the study of consumer choice, we pause briefly and use our interpretation of the slope of the indifference curve to discuss the third of the properties of the indifference curve—its characteristic curvature—which we left unexplained earlier. The shape of indifference curves means that the slope decreases with movement from left to right. In Figure 5-8, at point *m*, toward the right of the diagram, the consumer is willing to give up far less cheese for one more box of rubber bands (quantity *rm*) than she is willing to trade at point *M*, toward the left. This is because at *M* she initially has a large quantity of cheese and few rubber bands, while at *m* her initial stock of cheese is low and she has many rubber bands. In general terms, the curvature premise on which indifference curves are usually drawn asserts that consumers are relatively eager to trade away a commodity of which they have a large amount but are more reluctant to trade goods of which they hold small quantities. This psychological premise underlies the curvature of the indifference curve.

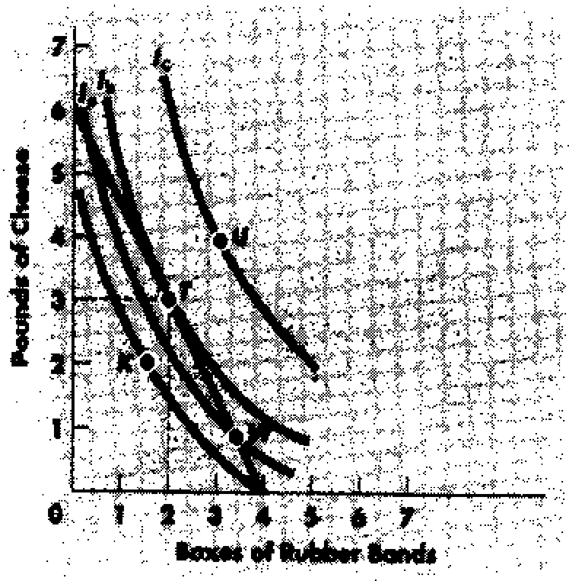
THE CONSUMER'S CHOICE

We can now use our indifference curve apparatus to analyze how the consumer chooses among the combinations that she can afford to buy, that is, the combinations of rubber bands and cheese shown by the budget line. Figure 5-9 brings together in the same diagram the budget line from Figure 5-4 and the indifference curves from Figure 5-7.

Since, according to the first of the properties of indifference curves, the consumer prefers higher curves to lower ones, she will go to the point on the budget line that lies on the highest indifference

FIGURE 5-9 OPTIMAL CONSUMER CHOICE

Point T is the combination of rubber bands and cheese that gives the consumer the greatest benefit for her money. I_b is the highest indifference curve that can be reached from the budget line. T is the point of tangency between the budget line and I_b.



curve attainable. This will be point T on indifference curve I_b. She can afford no other point that she likes as well. For example, neither point K below the budget line or point W on the budget line gets her on such a high indifference curve. Further, any point on an indifference curve above I_b, such as point U, is out of the question because it lies beyond her financial means. We end up with a simple rule of consumer choice:

Consumers will select the most desired combination of goods obtainable for their money. The choice will be that point on the budget line at which the budget line is tangent to an indifference curve.

We can see why no point except the point of tangency, T (two boxes of rubber bands and 3 pounds of cheese), will give the consumer the largest utility that her money can buy. Suppose that the consumer were instead to consider buying 3-1/2 boxes of rubber bands and one pound of cheese. This would put her at point W on the budget line and on the indifference curve I_a. But then, by buying fewer rubber bands and more cheese (a move to the left on the budget line), she could get to another indifference curve, I_b, that would be higher and hence more desirable without spending any more money. It clearly does not pay to end up at W. Only the point of tangency, T, leaves no room for improvement.

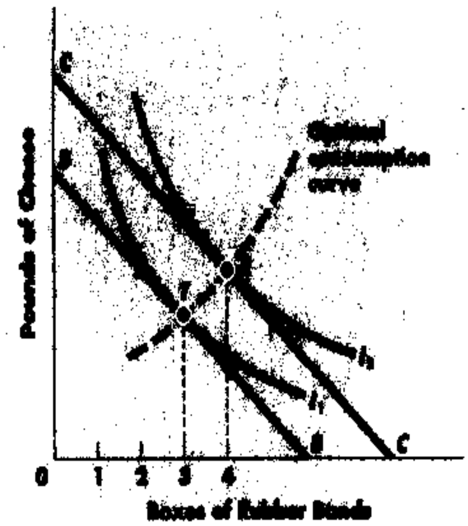
At a point of tangency, where the consumer's benefits from purchasing cheese and rubber bands are maximized, the slope of the budget line equals the slope of the indifference curve. This is true by the definition of a point of tangency. We have just seen that the slope of the indifference curve is the marginal rate of substitution between cheese and rubber bands, and that the slope of the budget line is the ratio of the prices of rubber bands and cheese. We can therefore restate the requirement for the optimal division of the consumer's money between the two commodities in slightly more technical language:

Consumers will get the most benefit from their money when they choose combinations of commodities whose marginal rates of substitution equal the ratios of their prices.

It is worth reviewing the logic behind this conclusion. Why is it not advisable for the consumer to stop at a point like W, where the marginal rate of substitution (slope of the indifference curve) is less than the price ratio (slope of the budget line)? Instead, by moving upward and to the left along her budget line, she can take advantage of market opportunities to obtain a commodity bundle that she likes better. This will always be true, for example, if the amount of cheese the consumer is *personally* willing to exchange for a box of rubber bands (the slope of the indifference curve) is greater

FIGURE 5-10 EFFECTS OF A RISE IN INCOME WHEN NEITHER GOOD IS INFERIOR

The rise in income causes a parallel shift in the budget line from *BB* to *CC*. The quantity of rubber bands demanded rises from three to four boxes, and the quantity demanded of cheese also increases.



than the amount of cheese for which the box trades *on the market* (the slope of the budget line).

CONSEQUENCES OF INCOME CHANGES: INFERIOR GOODS

Next, consider what happens to the consumer's purchases after a rise in income. We know that a rise in income produces a parallel outward shift in the budget line, such as the shift from *BB* to *CC* in Figure 5-10. This moves the consumer's equilibrium from tangency point *T* to tangency point *E* on a higher indifference curve.

A rise in income may or may not increase the demand for a commodity. In Figure 5-10, the rise in income does lead the consumer to buy more cheese *and* more rubber bands, but indifference curves need not always be positioned in a way that yields this sort of result. In Figure 5-11, as the consumer's budget line rises from *BB* to *CC*, the tangency point moves leftward from *H* to *G*, so that when her income rises she actually buys *fewer* rubber bands. This implies that rubber bands are an **inferior good.**

CONSEQUENCES OF PRICE CHANGES: DERIVING THE DEMAND CURVE

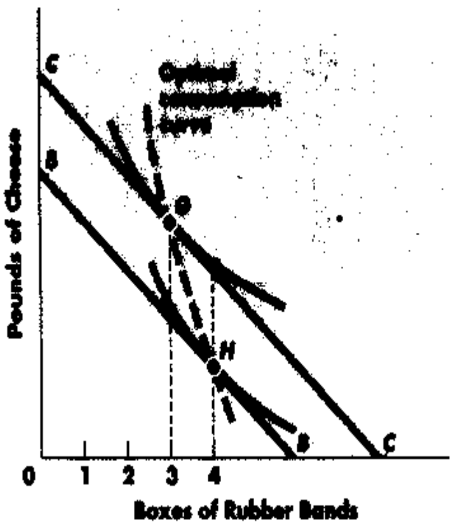
Finally, we come to the main question underlying demand curves: how does a consumer's choice change if the price of one good changes? We explained earlier that a reduction in the price of a box of rubber bands causes the budget line to swing outward along the horizontal axis while leaving its vertical intercept unchanged. In Figure 5-12, we depict the effect of a decline in the price of rubber bands on the quantity of rubber bands demanded. As the price of rubber bands falls, the budget line swings from *BC* to *BD*. The tangency points, *T* and *E*, also move in a corresponding direction, causing the quantity demanded to rise from two to three boxes. The price of rubber bands has fallen, and the quantity demanded has risen: the demand curve for rubber bands is negatively sloped.

The demand curve for rubber bands can be constructed directly from Figure 5-12. Point *T* shows that the consumer will buy two boxes of rubber bands when the price of a box is $3.00. Point *E* indicates that when the price falls to $1.50, quantity demanded rises to three boxes of rubber bands.[7] These two pieces of information are shown in Figure 5-13 as points *t* and *e* on the demand curve for rubber bands. By examining the effects of other

[7]How do we know that the price of rubber bands corresponding to the budget line *BD* is $1.50? Since the $12.00 total budget will purchase at most eight boxes (point *D*), the price per box must be $12.00/8 = $1.50.

FIGURE 5-11 EFFECTS OF A RISE IN INCOME WHEN RUBBER BANDS ARE AN INFERIOR GOOD

The upward shift in the budget line from *BB* to *CC* causes the quantity of rubber bands demanded to fall from four boxes (point *H*) to three (point *G*).



possible prices for rubber bands (other budget lines emanating from point *B* in Figure 5-12), we can find all the other points on the demand curve in exactly the same way.

The indifference curve diagram also brings out an important point that the demand curve does not show. A change in the *price of rubber bands* also has consequences for the *quantity of cheese demanded* because it affects the amount of money left over for cheese purchases. In the example illustrated in Figure 5-12, the decrease in the price of rubber bands increases the demand for cheese from 3 to 3 and 3/4 pounds.

FIGURE 5-12 CONSEQUENCES OF PRICE CHANGES

A fall in the rubber band price swings the budget line outward from *BC* to *BD*. The consumer's equilibrium point (the point of tangency between the budget line and an indifference curve) moves from *T* to *E*. The desired purchase of rubber bands increases from two to three boxes, and the desired purchase of cheese increases from 3 pounds to 3 and 3/4 pounds.

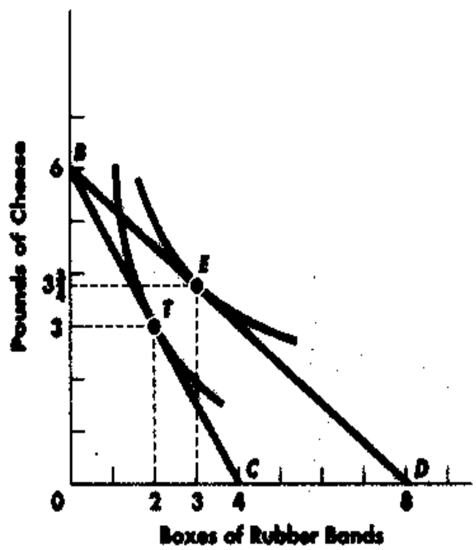
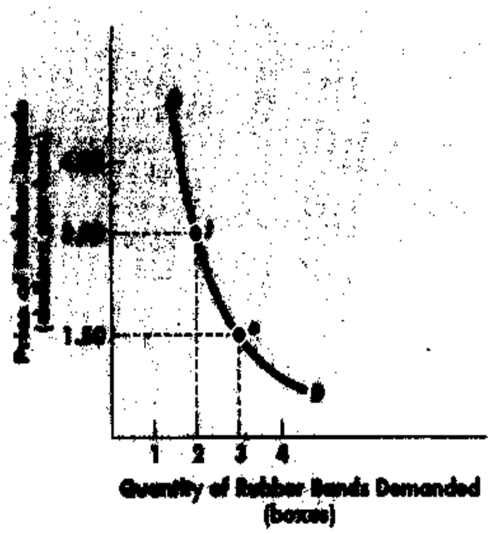


FIGURE 5-13 DERIVING THE DEMAND CURVE FOR RUBBER BANDS

The demand curve is derived from the indifference-curve diagram by varying the price of the commodity. Specifically, when the price is $3 per box of rubber bands, point *T* from Figure 5-12 indicates that the optimal purchase is two boxes. This information is recorded here as point *t*. Similarly, the optimal purchase is three boxes when the price of rubber bands is $1.50 (point *E* in Figure 5-12). This is shown here as point *e*.



SUMMARY

1. Indifference-curve analysis permits economists to study the interrelationships of the demands for two (or more) commodities.
2. The basic tools of indifference-curve analysis are the consumer's **budget line** and **indifference curves.**
3. A budget line shows all combinations of two commodities that the consumer can afford, given the prices of the commodities and the amount of money the consumer has available to spend.
4. The budget line is a straight line whose slope equals the ratio of the prices of the commodities. A change in price changes the slope of the budget line. A change in the consumer's income causes a parallel shift in the budget line.
5. Two points on an indifference curve represent two combinations of commodities such that the consumer does not prefer one combination over the other.
6. Indifference curves normally have negative slopes and are "bowed in" toward the origin. The **slope of an indifference curve** indicates how much of one commodity the consumer is willing to give up in order to get an additional unit of the other commodity.
7. The consumer will choose the point on her budget line that gets her to the highest attainable indifference curve. Normally this will occur at the point of tangency between the two curves. This point indicates the combination of commodities that gives her the greatest benefits for the amount of money she has available to spend.
8. The consumer's demand curve can be derived from her indifference curve.

KEY TERMS

Budget line
Indifference curve
Marginal rate of substitution
Slope of an indifference curve
Slope of a budget line

QUESTIONS FOR REVIEW

1. John Q. Public spends all of his income on gasoline and hot dogs. Draw his budget line under several conditions:
   a. His income is $80 and 1 gallon of gasoline and one hot dog each cost $1.60.
   b. His income is $120 and the two prices remain the same.
   c. His income is $80, hot dogs cost $1.60 each, and gasoline costs $1.20 per gallon.
2. Draw some hypothetical indifference curves for John Q. Public on a diagram identical to the one you constructed for the preceding question.
   a. Approximately how much gasoline and how many hot dogs will Mr. Public buy?
   b. How will these choices change if his income increases to $120, as in the previous question? Is either good an inferior good?
   c. How will these choices change if gasoline prices fall to $1.20 per gallon, as in the previous question?
3. Explain the information that the *slope* of an indifference curve conveys about a consumer's preferences. Use this relationship to explain the typical U-shaped curvature of indifference curves.

CHAPTER 6

DEMAND AND ELASTICITY

A high cross elasticity of demand [between two goods indicates that they] compete in the same market. [This can prevent a supplier of one of the products] from possessing monopoly power over price.

Supreme Court of the United States (The Du Pont Cellophane Decision, 1956)

Chapter 5 examined in some detail the origins of the individual's demand curve and showed how, by summing across individuals, economists derive the market demand curve. We explained that the market demand curve represents the relationship between price and the quantity demanded by all purchasers in the market, including both actual and prospective buyers. While the quantity demanded of any good depends on many things other than its price—such as the incomes of consumers and the prices of other goods—this chapter, just like the last, focuses on the *price* of a good.

We begin by examining certain critical properties of the demand curve—especially how to measure the responsiveness of quantity demanded to price. In particular, we introduce and explain an important concept called *elasticity* that economists almost always use for this purpose. Next, we apply that concept to measure the strength of the interconnection between the demands for related commodities, such as hamburgers and ketchup. Finally, an appendix explains the importance of the time period to which a demand curve applies and how this can create problems in obtaining demand information from statistical data.

TWO ILLUSTRATIVE CASES

Two examples—one hypothetical, the other concrete—will illustrate why decision makers in business and government need to measure the responsiveness of quantity demanded to price. As usual, we start by laying out the issues, postponing resolution until we have explained the required analytical tools.

EXAMPLE 1: THE REVENUE EFFECT OF A SALES-TAX CUT

Our first example is hypothetical, but it deals with a real issue that you will readily recognize. Leading politicians in the state of Confusion, painfully aware that pledging "no new taxes" is the key to reelection, decide to go that principle one better: They cut the state *sales tax* (called an *excise tax*) on beer and cigarettes by 20 percent, figuring that this will lead to a 20 percent cut in the tax revenues they obtain from each product. At the end of 2 years, however, they find to

their delight that revenues have fallen less than 20 percent. Furthermore, the decline in beer-tax revenues is much smaller than the decline in cigarette-tax revenues.

Why did they overestimate the revenue loss from cutting tax rates? Could they have anticipated that cigarette-tax revenues would fall more than beer-tax revenues? As we will see in this chapter, the answers hinge directly on the *responsiveness* of the quantities of beer and cigarettes demanded to changes in their prices.

EXAMPLE II: *POLAROID V. KODAK*

Our second example is about a multibillion-dollar court case involving two well-known U.S. firms in which one of the authors of this volume was deeply involved as an expert witness. Polaroid won its suit against Kodak for infringement of its instant-photography patents. A second court case was then launched to determine how much Kodak should be required to pay Polaroid in compensation. The issue centered on how much financial damage Polaroid had suffered as a result of Kodak's infringement.

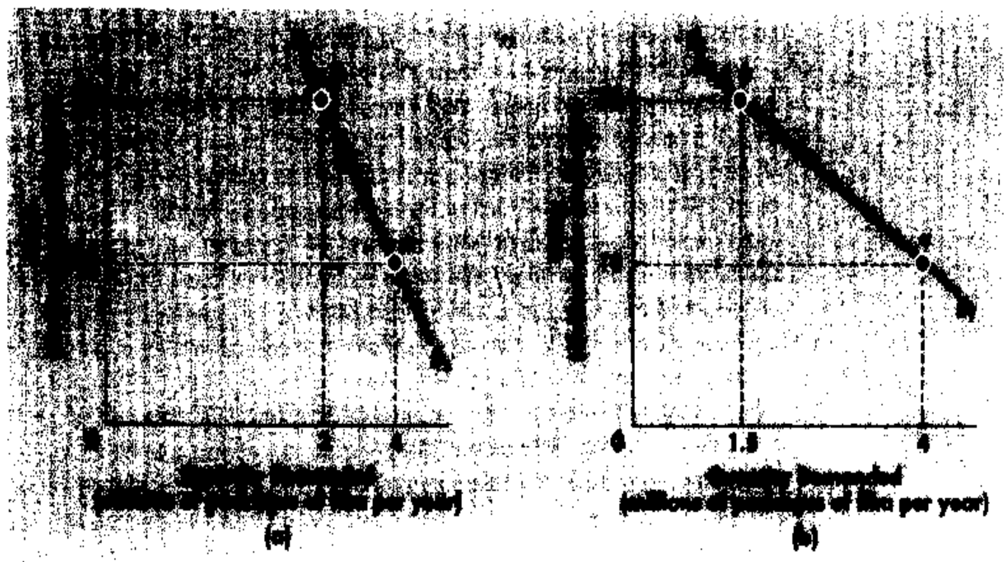
Polaroid claimed that it would have charged higher prices for its film and cameras if Kodak had not infringed its patents and had consequently been unable to compete with Polaroid. Witnesses for Kodak pointed out, however, that such price increases would have reduced the quantity of film and cameras demanded. The court was therefore faced with a clear question: How would Polaroid's total revenue have been affected by the price increases it claims it would have instituted?

To answer this question, recognize first that the outcome depends on the responsiveness of quantity demanded to price changes, that is, on the shape of the demand curve for Polaroid film. To illustrate the principles involved, we start with a hypothetical case in which demand is relatively unresponsive to price. In Panel (a) of Figure 6-1, a given (vertical) change in price leads to a small (horizontal) change in quantity, making the demand curve rather steep. With such a demand curve, if Polaroid had doubled its price from $10 per package of film to $20, the quantity sold would have fallen from 4 million to 3 million packages per year—the move from point *A* to point *B* on the steep demand curve in Figure 6-1(a).

According to testimony at the trial, Polaroid's management believed that such a price rise would increase the company's total sales revenue. The graph suggests that they were right. If the firm's demand curve was indeed very steep, a rise in price would have caused only a small drop in quantity sold. At the lower price of $10 per package, the firm could have sold 4 million packages per year, bringing total revenue of $10 times 4 million, or $40 million per year. At the higher price of $20 per package, however, it could have sold only 3 million packages annually. But these modestly lower sales would have yielded annual revenue of $20 times 3 million, or $60 million. Thus, in this example, doubling the price would have led to a 50 percent rise in Polaroid's total revenue.

Panel (b) of Figure 6-1 shows that things work out quite differently if quantity demanded is more responsive to price. Demand curve D_fD_f is much flatter than demand curve D_sD_s, meaning that a given change in price would elicit a much greater quantity response. As a result, if the price of film were to rise from $10 to $20, as it did in the previous graph, quantity demanded would have fallen far more sharply than before—from 4 million packages per year all the way down to 1.5 million. Consequently, the rise in price would have actually *reduced* Polaroid's total revenue from $40 million at point *a* down to $30 million ($20 × 1.5 million) at point *b*.

FIGURE 6-1 TWO HYPOTHETICAL DEMAND CURVES FOR POLAROID FILM



The demand curve in Panel (a) is comparatively steep, so that a rise in price leads to only a small change in quantity demanded. Consequently, the price increase raises Polaroid's total revenue. In contrast, the demand curve in Panel (b) is relatively flat, so that the same rise in price produces a sharp cut in quantity demanded and a drop in total revenue.

Thus, when the demand curve is relatively flat, as in Panel (b), a rise in price actually decreases total revenue. In this example, doubling the price would have produced a 25 percent drop in total revenue. The court had to decide whether the situation depicted in Panel (a) or that in Panel (b) was closer to the facts before it could resolve the dispute between Polaroid and Kodak.

Clearly, the *responsiveness* of quantity demanded to price was the key influence here. The court needed a good way to measure the responsiveness of quantity demanded to price changes—a subject to which we turn next.

ELASTICITY: THE MEASURE OF RESPONSIVENESS

It is not only courts and governments that need a way to measure the responsiveness of quantity demanded to price. Business firms do, too. They need it to make decisions on pricing of products, on whether to add new models of products, and so on. Economists measure this responsiveness by means of a concept they call *elasticity*. A demand curve like Figure 6-1(b), indicating that consumers respond sharply to a change in price, is said to be *elastic* or *highly elastic*. A demand curve like Figure 6-1(a), involving a relatively small or insignificant response by consumers to a given price change, is called *inelastic*.

The **(PRICE) ELASTICITY OF DEMAND** is the ratio of the *percentage* change in quantity demanded to the *percentage* change in price that brings about the change in quantity demanded.

The precise measure used for this purpose is called the **price elasticity of demand,** or sometimes simply the **elasticity of demand.** It is defined as the ratio of the *percentage* change in quantity demanded to the associated *percentage* change in price. Specifically:

$$\text{Elasticity of demand} = \frac{\text{Percentage change in quantity demanded}}{\text{Percentage change in price}}$$

Thus, demand is called *elastic* if, say, a 10 percent rise in price leads to a reduction in quantity demanded of more than 10 percent. Demand is called *inelastic* if such a rise in price reduces quantity demanded by less than 10 percent.

Let us now consider how these definitions can be used to analyze a particular demand curve. At first, it may seem that the *slope* of the demand curve conveys the needed information: Curve D_sD_s is much steeper than curve D_fD_f in Figure 6-1, so any given change in price appears to correspond to a much smaller change in quantity demanded in Figure 6-1(a) than in Figure 6-1(b). For this reason, it is tempting to call demand in Panel (b) "more elastic." But slope will not do the job because the slope of any curve depends on the units of measurement, as we showed in Chapter 2, and economists have no standardized units of measurement. Cloth output may be measured in yards or in meters, milk in quarts or liters, and coal in tons or kilograms.

Figure 6-2(a) brings out the point explicitly. In this graph, we return to the example of Chicken Maknoogat sales, measuring quantity demanded in terms of servings and price in dollars per serving. A fall in price from $1.70 to $1.60 per serving (points A and B) raises quantity demanded at one particular MakDoogal's restaurant from 300 servings to 380 servings per week.

Now look at Figure 6-2(b), which provides *exactly* the same information, but measures quantity demanded in ounces rather than servings. (One serving equals 10 ounces.) Thus, the same price change as before increases quantity demanded by 800 units, that is, by 800 ounces, rather than by 80 servings.

To the eye, the increase in quantity demanded looks 10 times as great in Panel (b) as in Panel (a), but all that has changed is the unit of measurement. The 800-unit increase in Figure 6-2(b) represents the same increase in quantity demanded as the 80-unit increase in Figure 6-2(a). Just as you get different numbers for a given rise in temperature depending on whether you measure it in Celsius or Fahrenheit, the slope of the demand curve is different depending on whether you measure quantity in servings or in ounces. Clearly, then, slope is not a good measure of the responsiveness of quantity demanded to price, because the measure changes whenever the units of measurement change.

It is because of this problem that economists use the elasticity measure, which is based on *percentage* changes in price and quantity rather than on *absolute* changes. The elasticity formula solves the units problem because percentages are unaffected by units of measurement. If the defense budget doubles, it goes up by 100 percent, whether measured in millions or billions of dollars. If a person's height doubles between the ages of 2 and 18, it goes up 100 percent, whether measured in inches or centimeters. The elasticity formula given above expresses both the change in quantity demanded and the change in price as *percentages*.[1]

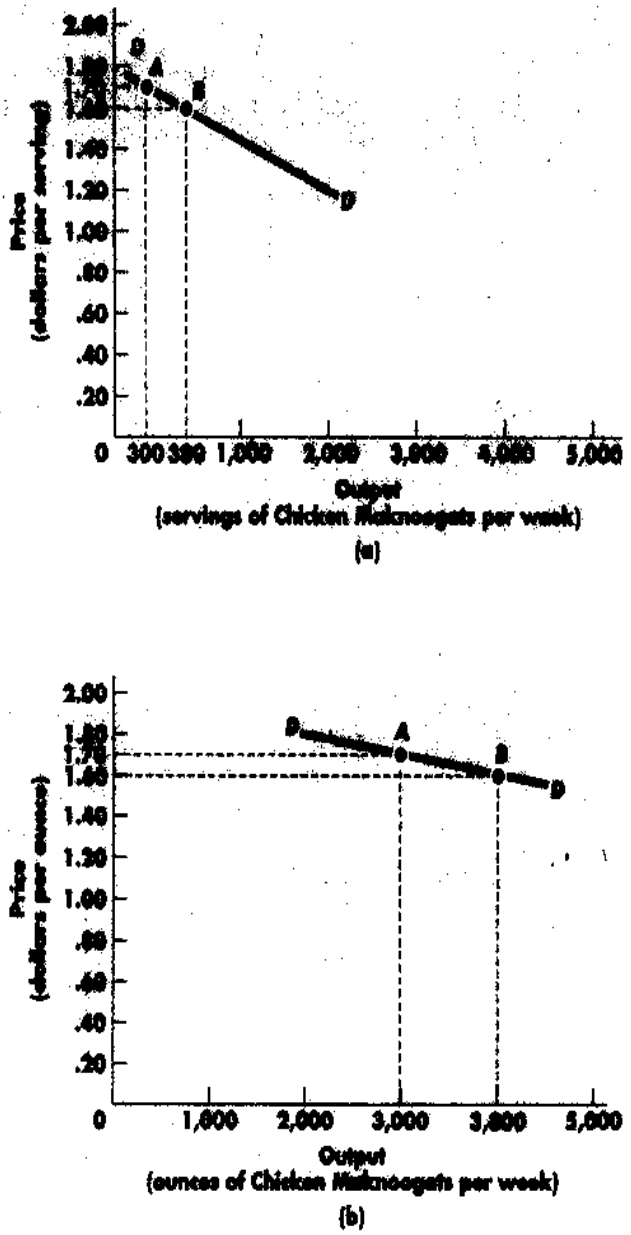
In addition to calculating elasticity in terms of percentages, the formula usually used in practice has a second important attribute: the change in quantity is calculated neither as a percentage of the "initial" quantity nor as a percentage of the "subsequent" quantity, but *as a percentage of the average of the two quantities*. Similarly, the change in price is expressed as a percentage of the average of the two prices. To understand why, consider, as an example, the demand information presented in Figure 6-1(a). At a price of $20, quantity demanded equaled 3 million packages of film; at a price of $10, quantity demanded equaled 4 million. The difference in sales volume is 4 million minus 3 million, or 1 million

---

[1]The remainder of this section involves fairly technical computational issues. On a first reading you may prefer to go directly to the new section that begins on page 130.

FIGURE 6-2 THE SENSITIVITY OF SLOPE TO UNITS OF MEASUREMENT

The slope of a curve changes whenever units of *measurement change*. Panel (a)'s demand curve looks quite steep because it measures output in servings of Maknoogats. Panel (b) measures quantity in ounces instead. With 10 ounces per serving, all of the quantities are multiplied by 10. As a result, the demand curve looks rather flat, although the two demand curves present exactly the same information.



packages, which is 33.3 percent of 3 million, but 25 percent of 4 million. Which is the correct figure to use as the percentage change in quantity?

This dilemma always occurs because any given change in quantity must involve some larger quantity Q (4 million in the film example) and some smaller quantity q (3 million), so that the same change in quantity must be a smaller percentage of Q and a larger percentage of q. Obviously, neither of these can claim to be *the* right percentage change in quantity. It turns out to be convenient to use what appears to be a compromise—the *average* of the two quantities. In the film example, we express the 1 million package change in

quantity as a percentage of the average of 4 million and 3 million, or 3.5 million. Thus:

$$\text{Percentage change in quantity} = 1 \text{ million as a percentage of } 3.5 \text{ million} = 28.6 \text{ percent}$$

This is a kind of compromise between 33.3 percent and 25 percent. Similarly, in calculating the percentage change in price, take the $10 change in price as a percentage of the average of $10 and $20, giving $10/$15, or 66.7 percent.

To summarize, the elasticity formula has two basic attributes:

1. Each of the changes with which it deals is measured as a *percentage* change.
2. Each of the percentage changes is calculated in terms of the average values of the quantities and prices.

In addition, the formula for price elasticity of demand usually is adjusted in a third way. Note that when the price increases, the quantity demanded usually declines. This results in a negative ratio of the two percentage changes. However, it is customary to express elasticity as a positive number. Hence:

3. Each percentage change is taken as an "absolute value," meaning that the calculation drops all minus signs.[2]

We can now restate the formula for price elasticity of demand, keeping in mind all three features of the formula:

$$\text{Price elasticity of demand} = \frac{\text{Change in quantity as a percentage of the average of the two quantities}}{\text{Change in price as a percentage of the average of the two prices}}$$

In the Polaroid film example:

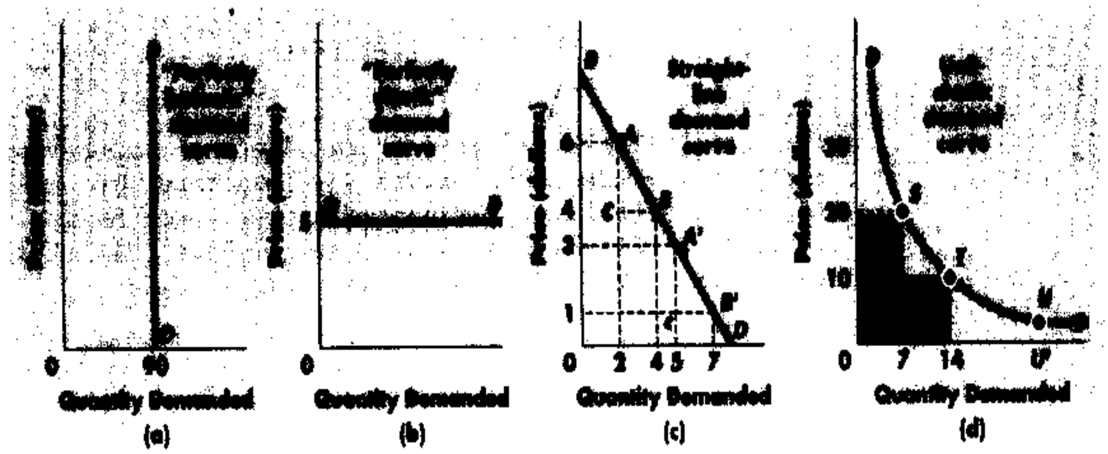
$$\text{Elasticity} = \frac{1 \text{ million as a percentage of } 3.5 \text{ million}}{\$10 \text{ as a percentage of } \$15} = \frac{28.6 \text{ percent}}{66.7 \text{ percent}} = 0.43$$

PRICE ELASTICITY OF DEMAND AND THE SHAPES OF DEMAND CURVES

We noted earlier that looks can be deceiving because of the arbitrariness of units of measure. That is why the elasticity formula was invented. Nonetheless, the shape of a demand curve does give some information about its elasticity. Let us see what that is, with the aid of Figure 6-3.

[2]This third attribute of the elasticity formula—the removal of all minus signs—applies only when the formula is used to measure the responsiveness of *quantity demanded* to a change in *price*. Later in the chapter, we will show that similar formulas are used to measure the responsiveness between other pairs of variables. For example, the elasticity of supply uses a similar formula to measure the responsiveness of quantity supplied to price. In such other uses, it is not customary to drop minus signs when calculating elasticity. The reasons will become clearer later in the chapter.

FIGURE 6-3 DEMAND CURVES WITH DIFFERENT ELASTICITIES



The vertical demand curve in Panel (a) is *perfectly inelastic* (elasticity = 0), so quantity demanded remains the same regardless of price. The horizontal demand curve in Panel (b) is *perfectly elastic*—at any price above $5, quantity demanded falls to zero. Panel (c) shows a *straight-line demand curve*. Its *slope* is constant, but its *elasticity* is not. Panel (d) depicts a *unit-elastic* demand curve whose constant elasticity is 1 throughout. A change in price pushes quantity demanded in the opposite direction but does not affect total expenditure. When price equals $20, total expenditure is price times quantity, or $20 × 7 = $140; when price equals $10, expenditure equals $10 × 14 = $140.

We begin with two extreme but important cases. Panel (a) depicts a demand curve that is simply a vertical line. Such a curve is called *perfectly inelastic* throughout because its elasticity is zero. Since quantity demanded remains at 90 units no matter what the price, the percentage change in quantity is always zero, and hence the elasticity is always zero. Thus, in this case, consumer purchases do not respond at all to any change in price.

Vertical demand curves like this are quite unusual. They sometimes occur for products with particularly low prices and that are normally used with something else. For example, will anyone use more salt if its price falls? The demand curve may also be vertical when the item in question is considered an absolute necessity by the consumer (something like an essential medicine), although even here, the demand curve will remain vertical only so long as price does not exceed what the consumer can afford.

Panel (b) of Figure 6-3 shows the opposite extreme: a horizontal demand curve. This curve is said to be *perfectly elastic* (or *infinitely elastic*). At the slightest rise in price, quantity demanded will drop to zero; that is, the percentage change in quantity demanded will be infinitely large. This can occur when a rival product that is just as good in the consumer's view is available at the going price ($5 in the diagram). In cases where no one will pay more than the going price, the seller will lose all of her customers if she raises her price by even a penny.

Panel (c) depicts a case between these two extremes: a *straight-line* demand curve that runs neither vertically nor horizontally. Though the *slope* of a straight-line demand curve remains constant throughout its length, its *elasticity* does not. For example, the elasticity of demand between points *A* and *B* in Figure 6-3(c) is:

$$\frac{\text{Change in } Q \text{ as a percentage of average } Q}{\text{Change in } P \text{ as a percentage of average } P} = \frac{2/3}{2/5} = \frac{66.67 \text{ percent}}{40 \text{ percent}} = 1.67$$

But the elasticity of demand between points A' and B' is:

$$\frac{2 \text{ as a percentage of } 6}{2 \text{ as a percentage of } 2} = \frac{33.33 \text{ percent}}{100 \text{ percent}} = 0.33$$

The general point is that:

Along a straight-line demand curve, the price elasticity of demand grows steadily smaller with movement from left to right. That is because the quantity keeps getting larger, so that a given *numerical* change in quantity becomes an ever smaller *percentage* change. But the price keeps going lower, so that a given numerical change in price becomes an ever larger percentage change.

If the elasticity of a straight-line demand curve varies from one part of the curve to another, what does a demand curve with the same elasticity throughout its length look like? For reasons explained in the next section, it has the general shape indicated in Figure 6-3(d). That panel shows a curve with elasticity equal to 1 throughout (a *unit-elastic* demand curve). A unit-elastic demand curve bends in the middle toward the origin of the graph; at either end, it moves closer and closer to the axes but never touches or crosses them.

As we have noted, a curve with an elasticity greater than 1 is called an *elastic demand curve* (one for which the percentage change in quantity is greater than the percentage change in price); a curve whose elasticity is less than 1 is known as an *inelastic* curve. When elasticity is exactly 1, economists say that the curve is *unit elastic*.

Real-world price elasticities of demand seem to vary considerably from product to product. Moderate luxury goods such as expensive vacations are generally more price-elastic than goods like milk and bread that are considered necessities. Products with close substitutes tend to have relatively high elasticities. Also, the elasticities of demand for producers' goods such as raw materials and machinery tend to be greater on the whole than those for consumers' goods. Table 6-1 gives actual statistical estimates of elasticities for some commodities in the economy.

THE RELATIONSHIP BETWEEN ELASTICITY AND SLOPE

We have emphasized that elasticity and slope are not the same thing. Still, the two are closely related, and this relationship helps to explain the shapes of the curves in Figure 6-3. Our formula for elasticity can be rewritten more compactly in symbols:

$$\text{Price elasticity of demand} = \frac{\Delta Q/Q}{\Delta P/P}$$

where ΔQ and ΔP are the changes in quantity and price, respectively, P is the average price, and Q is the average quantity. This version of the formula just rewrites the definition of price elasticity of demand as the percentage change in Q divided by the percentage change in P, except in one minor respect: it replaces percentages (like 25 percent) with decimal fractions (like 0.25).

This formula can also usefully be rewritten as:

$$\text{Price elasticity of demand} = \frac{\Delta Q/\Delta P}{Q/P}$$

TABLE 6-1

ESTIMATES OF LONG-RUN PRICE ELASTICITY OF DEMAND: 1947–1964

Commodity	Long-Run Price Elasticity
Aluminum	0.4
Shoe repairs and cleaning	0.4
Newspapers and magazines	0.5
Medical care and hospitalization insurance	0.8
Purchased meals (excluding alcoholic beverages)	1.6
Electricity (household utility)	1.9
Boats, pleasure aircraft	2.4
Public transportation	3.5
China, tableware	8.8

SOURCE: The aluminum figure is from Franklin M. Fisher, *A Priori Information and Times Series Analysis, Essays in Economic Theory and Management* (Amsterdam: North Holland Publishing Company, 1962), p. 112. All other figures are from H. S. Houthakker and Lester D. Taylor, *Consumer Demand in the United States*, 2nd ed. (Cambridge, Mass.: Harvard University Press, 1970), pp. 153–158.

where $\Delta Q/\Delta P$ is the reciprocal of the slope of the demand curve, which is $\Delta P/\Delta Q$. Clearly, slope is part of the elasticity formula. But it is only part: the ratio Q/P is involved, as well.

This last formula indicates why a vertical demand curve, like Figure 6-3(a), has an elasticity of zero. When the demand curve is vertical, no change in price ever changes quantity at all, so ΔQ is always zero. Since anything multiplied by zero equals zero, the last formula confirms that the elasticity of a vertical demand curve is always zero. Similarly, if the demand curve is horizontal, as in Figure 6-3(b), price never changes, so P equals zero. Since any number divided by zero is, roughly speaking, infinity, a horizontal demand curve has infinite elasticity.

The varying elasticity of the straight-line demand curve in Figure 6-3(c) is also explained by this last version of the elasticity formula. Along a straight line, slope is the same everywhere, by definition, so $\Delta Q/\Delta P$ must be constant. But, as we move from one part of the straight-line demand curve to another, the ratio Q/P changes constantly. Toward the left-hand end of the demand curve, P is relatively large and Q is relatively small, so the ratio Q/P is large. The formula thus calls for a high elasticity. The opposite is clearly true toward the right-hand end.

PRICE ELASTICITY OF DEMAND AND TOTAL EXPENDITURE

Aside from its role as a measure of the responsiveness of demand to a change in price, elasticity serves a second purpose. Often, as in the Polaroid example at the beginning of this chapter, a firm wants to know whether a rise in price will increase or decrease its revenue. The price elasticity of demand provides a simple guide to the answer:

If demand for the seller's product is elastic, a rise in price will decrease total revenue. If demand is exactly unit elastic, a rise in price will leave total revenue

unaffected. If demand is inelastic, a rise in price will raise total revenue. The opposite changes will occur when price falls.

Note that a corresponding story must hold true about the expenditures made by the *buyers* of the product. After all, the expenditures of the buyers are exactly the same thing as the revenues of the seller.

These relationships between elasticity and total revenue hold because total revenue (or expenditure) equals price times quantity demanded, $P \times Q$, and a fall in price has two opposing effects on that arithmetic product. It decreases P, and, if the demand curve is negatively sloped, it increases Q. The first effect decreases revenues by cutting the amount of money that consumers spend on each unit of the good. But the second effect increases revenues by raising the number of units of the good that the firm sells.

The net effect on total revenue (or total expenditure) depends on the elasticity. If price goes down 10 percent and quantity demanded increases 10 percent (a case of *unit elasticity*), the two effects just cancel out: $P \times Q$ remains constant. On the other hand, if price goes down 10 percent and quantity demanded rises 15 percent (a case of *elastic* demand), $P \times Q$ increases. Finally, if a 10 percent price fall leads to a 5 percent rise in quantity demanded (*inelastic* demand), $P \times Q$ falls.

The relationship between elasticity and total revenue is easily seen in a graph. First, note that:

The total revenue (or expenditure) represented by any point on a demand curve (any price-quantity combination), such as point S in Figure 6-4, equals the area of the rectangle under that point (the area of rectangle $ORST$ in the figure). This is because the area of a rectangle equals its height times width, OR times RS. This is, clearly, price times quantity, which is exactly total revenue.

To illustrate the connection between elasticity and consumer expenditure, Figure 6-4 shows an elastic portion of a demand curve, DD. At a price of \$6 per unit, the quantity sold is four units, so total expenditure is $4 \times \$6 = \$24$, which is represented by the vertical rectangle whose upper right-hand corner is point S. When price falls to \$5 per unit, 12 units are sold. Consequently, the new expenditure ($\$60 = \$5 \times 12$), measured by the rectangle OUVW, exceeds the old.

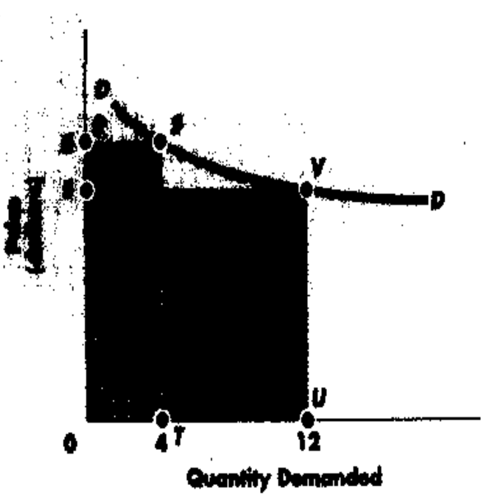
In contrast, Figure 6-3(d), the unit-elastic demand curve, shows constant expenditures even though price changes. Total spending is \$140 whether the price is \$20 and 7 units are sold (point S) or the price is \$10 and 14 units are sold (point T).

This discussion also indicates why a unit-elastic demand curve must have the shape depicted in Figure 6-3(d), hugging the axes closer and closer but never touching or crossing them. We have seen that when demand is unit elastic, total expenditure must be the same at every point on the curve. It must equal \$140 at point S and point T and point U. Suppose that at point U (or some other point), the demand curve were to touch the horizontal axis, meaning that the price would equal zero. Then total expenditure would be zero, not \$140. Therefore, if the demand curve remains unit elastic all along its length, it can never cross the horizontal axis (where $P = 0$). By the same reasoning, it cannot cross the vertical axis (where $Q = 0$). Since the slope of the demand curve is negative, the curve simply must get closer and closer to the axes with movement away from its middle points, as illustrated in Figure 6-3(d).

All of this indicates why elasticity of demand is so important for business decisions. A firm should not jump to the conclusion that an increase in price will

FIGURE 6-4 AN ELASTIC DEMAND CURVE

When price falls, quantity demanded rises by a greater percentage, increasing the total expenditure. Thus, when the price falls from $6 to $5, quantity demanded rises from 4 to 12 units, and total expenditure rises from $6 × 4 = $24 to $5 × 12 = $60.



add to its profits, for it may find that consumers take their revenge by cutting back on their purchases. For example, when Heinz's Pet Products Co. raised the prices of its 9-Lives cat food a few years ago, its market share plummeted from 23 percent to 15 percent as customers flocked to other brands. In fact, if the demand curve is elastic, a firm that raises price will end up selling so many fewer units that its total revenue will actually fall, even though it makes more money than before on each unit it sells.

Price cuts can also be hazardous—if the elasticity of demand is low. For example, cigarettes have an estimated price elasticity of approximately 0.4, meaning that a 10 percent drop in price can be expected to induce only a 4 percent rise in demand. This may explain why, when Philip Morris cut the price of Marlboros by about 18 percent, the company's profits dropped by 25 percent within months. In sum, the strategic value to a business firm of a price rise or a price cut depends very much on the elasticity of demand for its product. As the box on page 136 shows, a poor estimate of elasticity for a major product can be a devastating event for a company.

WHAT DETERMINES ELASTICITY OF DEMAND?

What kinds of goods have elastic demand curves, meaning that quantity demanded responds strongly to price? What kinds of goods have inelastic demand curves? Several considerations are relevant.

NATURE OF THE GOOD

Necessities, such as basic foodstuffs, normally have very inelastic demand curves, meaning that their quantities demanded are not very responsive to changes in their prices. For example, the quantity of potatoes demanded does not decline much when the price of potatoes rises. One study estimated that the price elasticity of demand for potatoes is just 0.3, meaning that when their price rises 10 percent, the quantity of potatoes purchased falls only 3 percent. In contrast, many *luxury goods*, such as restaurant meals, have rather elastic demand curves. One estimate found that the price elasticity of demand for restaurant meals is 1.6, so that a 10 percent price rise will cut purchases by 16 percent.

PRICE ELASTICITY: IGNORE IT AT YOUR PERIL



Apple Computer learned the hard way the importance of a good estimate of price elasticity when it put a relatively high price tag on its Macintosh personal computer. Its error not only reduced revenues for Apple but also ultimately cost the founder, Steve Jobs, and his successor, John Scully, their positions with the firm.

In 1983, Apple was preparing to introduce its first Macintosh to the market. Having just endured a brutal price war over its Apple II product line, management was eager to put price competition behind them. They decided that the Mac's unique features and novel, easy-to-use software could support a higher price than rival models. While the firm's original goal had been to design a computer that would sell for about $1,000, the final product hit the shelves priced at $2,500.

Consumers proved unwilling to pay a premium for the Mac. After an initial spurt, sales plunged. Large companies shunned the Mac, opting to stick with IBM computers. By 1990, Apple's market share had become so small software developers did not bother developing new products for it. Worse, Microsoft came out with new Windows 3.0 software that offered IBM computers many of the user-friendly features that were the hallmark of the Mac.

Apple finally relented, introducing new low-cost Macs and slashing prices on its existing products. Using elasticity measures, Apple's management predicted that the price cuts would increase sales on all models. Instead, the new sales were concentrated at the bottom end of the product range.

Apple's problems led to cutbacks and layoffs and then to the ouster of Scully, who was replaced by Michael Spindler in 1993. Spindler is keeping a closer eye on demand to determine what is selling and what isn't. He also has overseen the launch of the firm's new Power Macs, the first Macs in years that can compete on both price and performance. Armed with the hard-won knowledge it has acquired over the last ten years, Apple is betting that it can finally come up with the right product at the right price.

SOURCES: "Spindler's Apple," *Business Week*, October 3, 1994, pp. 88–96; "An Apple a Day," *The Economist*, August 24, 1991, pp. 61–62.

AVAILABILITY OF CLOSE SUBSTITUTES

If consumers can easily obtain a good substitute for a product whose price increases, they will switch readily. Thus, when the market offers close substitutes for a given product, its demand will be more elastic. This is sometimes a critical determinant of elasticity. The demand for gasoline is inelastic because it is not easy to run a car without fuel, but the demand for *any particular brand* of gasoline is extremely elastic, because other brands will work just as well. This example suggests a general principle: the demand for narrowly defined commodities (such as iceberg lettuce) is more elastic than the demand for more broadly defined commodities (such as vegetables).

FRACTION OF INCOME ABSORBED

The fraction of income absorbed by a particular item is also important. Very inexpensive items tend to have inelastic demands. Who will buy fewer shoelaces if the price of shoelaces rises? But families will buy fewer cars if auto prices go up.

PASSAGE OF TIME

This factor is relevant because the demand for many products is more elastic in the long run than in the short run. For example, when the price of home heating oil rose in the 1970s, some homeowners switched from oil heat to gas heat.

Very few of them switched immediately, however, so the short-term demand for oil was quite inelastic. As time passed and more homeowners had the opportunity to purchase and install new equipment, however, the demand curve gradually became more elastic.

We will show in the appendix to this chapter that the price elasticity of demand is not easy to calculate statistically. First, though, we will give some examples of important elasticity measures other than *price* elasticity of demand.

ELASTICITY AS A GENERAL CONCEPT

While we have spent much time studying the *price* elasticity of demand, we cannot ignore the role of elasticity as a general measure of the responsiveness of one economic variable to another. It is clear from our earlier discussion that a firm will be keenly interested in the price elasticity of its demand curve, but its interest in demand does not end there. As we have noted, quantity demanded depends on other things besides price. Business firms will be interested in consumer responsiveness to changes in these variables, as well.

For example, quantity demanded depends on consumer incomes. The firm's managers will, therefore, want to know how much a change in consumer income will affect the quantity of its product demanded. Fortunately, an elasticity measure can be helpful here, too. An increase in consumer incomes clearly raises the quantity demanded of most goods. To measure the response, economists use the *income elasticity of demand*—defined as the ratio of the percentage change in quantity demanded to the percentage change in income.

They also use elasticity to measure other, analogous responses. For example, to measure the response of quantity *supplied* to a change in price, economists use the *price elasticity of supply*, defined as the ratio of the percentage change in quantity supplied to the percentage change in price. The logic and analysis of all such elasticity concepts are, of course, perfectly analogous to those for price elasticity of demand.

CROSS ELASTICITY OF DEMAND: SUBSTITUTES AND COMPLEMENTS

Two goods are called **COMPLEMENTS** if an increase in the quantity consumed of one increases the quantity demanded of the other, all other things remaining constant.

Two goods are called **SUBSTITUTES** if an increase in the quantity consumed of one cuts the quantity demanded of the other, all other things remaining constant.

For many products, the quantities demanded depend on the quantities and prices of other products. Certain goods make others more desirable. For example, cream and sugar increase the desirability of coffee, and vice versa. The same is true of mustard or ketchup and hamburgers. In some extreme cases, neither product ordinarily has any use without the other—automobiles and tires, shoes and shoelaces, and so on. Such goods, each of which makes the other more valuable, are called **complements.**

The demand curves of complements are interrelated; specifically, a rise in the price of coffee is likely to reduce the quantity of sugar demanded. Why? When coffee prices rise, people drink less coffee and therefore demand less sugar to sweeten it. The opposite will be true of a fall in coffee prices. A similar relationship holds for other complementary goods.

At the other extreme, some goods make one another *less* valuable. These are called **substitutes.** Ownership of a motorcycle, for example, may decrease one's desire for a bicycle. If your pantry is stocked with cans of tuna fish, you are less likely to rush out and buy cans of salmon. As you might expect, demand curves for substitutes are also interrelated, but in the opposite direction. When the price

of motorcycles falls, people may demand fewer bicycles, so the quantity demanded falls. When the price of salmon goes up, people may eat more tuna.

The **CROSS ELASTICITY OF DEMAND** for Product *X* to a change in the price of another product, *Y*, is the ratio of the percentage change in quantity demanded of *X* to the percentage change in the price of *Y* that brings about the change in quantity demanded.

Economists can use another elasticity measure to determine whether two products are substitutes or complements: their **cross elasticity of demand.** This measure is defined much like the ordinary price elasticity of demand, only instead of measuring the responsiveness of the quantity demanded of, say, coffee to a change in its own price, cross elasticity of demand measures the responsiveness of the quantity demanded of coffee to a change in the price of, say, sugar. For example, if a 20 percent rise in the price of sugar reduces the quantity of coffee demanded by 5 percent (a change of *minus* 5 percent in quantity demanded), then the cross elasticity of demand will be

$$\frac{\text{Percentage change in quantity of coffee demanded}}{\text{Percentage change in sugar price}} = \frac{-5\%}{20\%} = -0.25$$

Obviously, cross elasticity is important for business firms, especially where the prices charged by their rivals are concerned. Northwest Airlines, for example, knows all too well that it will lose customers if it does not match price cuts by Delta or United. Coke and Pepsi provide another clear case in which cross elasticity of demand is crucial. But interest in cross elasticity is not limited to direct competitors. For example, the prices of VCRs and video rentals may have profound effects on the quantity of theater tickets demanded.

The cross elasticity of demand measure underlies the following rule about complements and substitutes:

If two goods are substitutes, a rise in the price of one of them tends to raise the quantity demanded of the other; so their cross elasticities of demand will normally be positive. If two goods are complements, a rise in the price of one of them tends to decrease the quantity demanded of the other item, so their cross elasticities will normally be negative.[3]

This result is really a matter of common sense. If the price of a good rises and buyers can find a substitute, they will tend to switch to the substitute. If the price of Japanese cameras goes up and the price of American cameras does not, at least some people will switch to the American product. Thus, a *rise* in the price of Japanese cameras causes a *rise* in the quantity of American cameras demanded. Both percentage changes are positive numbers and so their ratio, the cross elasticity of demand, is also positive.

On the other hand, if two goods are complements, a rise in the price of one will discourage both its own use and use of the complementary good. Automobiles and car radios are obviously complements. A large increase in the price of cars will depress sales of cars, and this in turn will reduce sales of car radios. Thus, a positive percentage change in the price of cars leads to a negative percentage change in the quantity of car radios demanded. The ratio of these numbers, the cross elasticity of demand for cars and radios, is therefore negative.

In practice, courts of law often evaluate cross elasticity of demand to measure whether particular firms face strong competition that can prevent them from overcharging consumers. If a rise in the price of Firm X causes consumers of its product to switch in droves to a competitor's product, Y, then the cross elasticity of demand for Product Y with respect to the price of X is high. That, in turn,

[3]Because cross elasticities can be positive or negative, it is *not* customary to drop minus signs as in a calculation of the ordinary price elasticity of demand.

HOW LARGE IS A FIRM'S MARKET SHARE? CROSS ELASTICITY AS A TEST



A firm's "market share" is often a crucial element in antitrust lawsuits (see Chapter 19), and for a simple reason. If the firm supplies no more than, say, 20 percent of the industry's output, courts and regulators presume that the firm is not a monopoly, since its competitors account for most of the industry's sales. On the other hand, if the defendant firm in the lawsuit accounts for 90 percent of the industry's output, courts may have good reason to worry about monopolization.

For this reason, court cases often feature lively debates in which the defendant firms try to prove that they have very small market shares while the plaintiffs seek to establish the opposite. Each side knows how much the defendant firm actually produces and sells, so what do they find to argue about? The dispute is about *the size of the total relevant market*. Ambiguity arises here because different firms do not produce identical products. For instance, are Rice Krispies in the same market as Cheerios? And how about Quaker Oatmeal, which users eat hot? What about frozen waffles? Are all of these products part of the same market?

Many observers argue, as the Supreme Court did in the famous Du Pont cellophane case (and, more recently, when Northwest Airlines and Continental Airlines sued American Airlines in 1993) that one of the proper criteria for determining the size of the relevant market is *cross elasticity of demand*. If two products have a high and positive cross elasticity, they must be close substitutes that compete closely, so they must be in the same market. But how large must the cross elasticity be before the court decides that two products are in the same market? While the law has established no clear elasticity number to determine whether a particular firm is in a relevant market, a number of courts of law have determined that a very high cross elasticity number clearly indicates effective competition.

means that competition is really powerful enough to prevent Firm X from raising its price arbitrarily. This is why cross elasticity is used so often in litigation before courts or government regulatory agencies when the degree of competition is an important issue (see the box above).

SHIFTS OF THE DEMAND CURVE: ADVERTISING, INCOME, AND THE PRICES OF COMPLEMENTS AND SUBSTITUTES

Demand is obviously a complex phenomenon. We have studied in detail the dependence of quantity demanded on price, and we have noted that quantity demanded depends on other variables such as consumers' incomes and the prices of complementary and substitute products. Because of changes in these other variables, demand curves often do not keep the same shape and position as time passes. Instead, they shift about. Chapter 4 showed that shifts in demand curves have predictable consequences for both quantity and price.

Public policy and business discussions often make vague references to a "change in demand." By itself, this expression does not mean anything. Remember from our discussion in Chapter 4 that it is vital to distinguish between a response to a price change—*which shows up as a movement along the demand curve*—and a change in the relationship between price and quantity demanded—*which produces a shift in the demand curve*—in effect, it changes the curve itself.

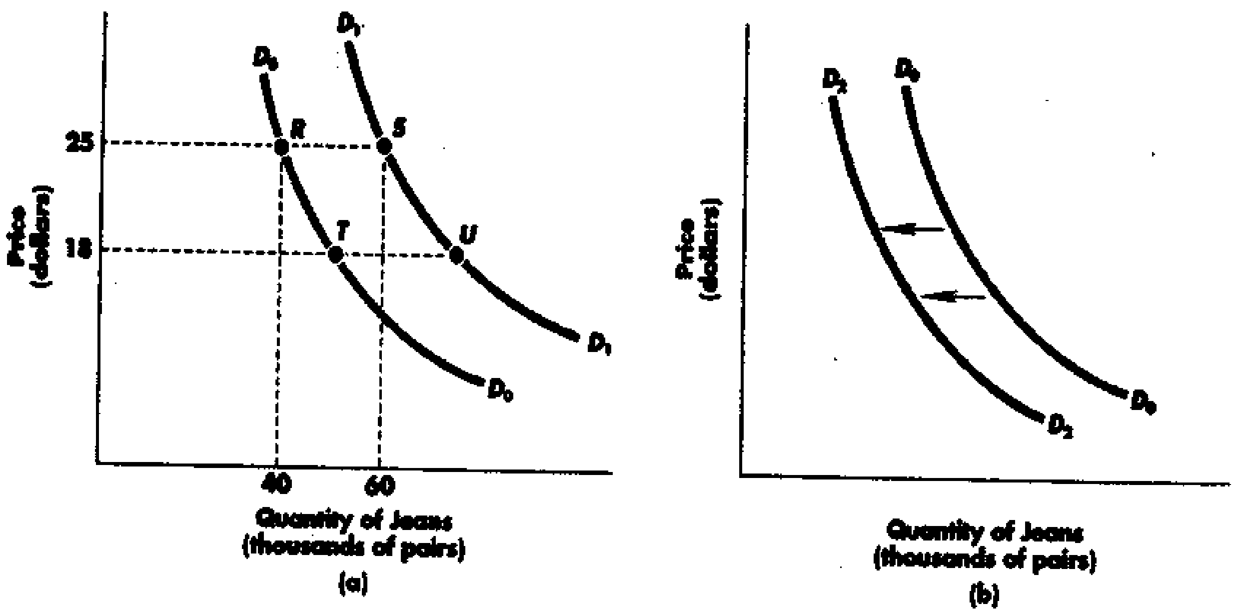
When price falls, quantity demanded generally responds by rising. This is a movement *along* the demand curve. On the other hand, an effective advertising campaign may mean that people will buy more goods at *any given price*. This would cause a rightward *shift* of the demand curve. Such a shift can be caused by a change in the value of any of the variables other than price that affect quantity demanded. While the distinction between a shift in a demand curve and a

movement along it may at first seem trivial, it is a significant difference in practice and can cause confusion if it is ignored. Let us pause for a moment to consider how changes in some of these other variables shift the demand curve.

As an example, consider the effect of a change in consumer income on the demand curve for jeans. In Figure 6-5(a), the black curve D_0D_0 is the original demand curve for jeans. Now suppose that parents start sending more money to their needy sons and daughters in college. If the price of jeans were to stay the same, those students are likely to use some of their increased income to buy more jeans. For example, if the price were to remain at \$25, quantity demanded might rise from 40,000 pairs (point *R*) to 60,000 (point *S*). Similarly, if price had instead been \$18, and had remained at that level, the rise in income might produce a corresponding change from *T* to *U*. In other words, the rise in income would be expected to *shift* the entire demand curve to the right from D_0D_0 to D_1D_1. In exactly the same way, a fall in consumer income can be expected to lead to a leftward shift in the demand curve for jeans, as shown in Figure 6-5(b).

Other variables that affect quantity demanded can be analyzed in the same way. For example, an increase in TV advertising for jeans might lead to a rightward (outward) shift in the demand curve for jeans, as in Figure 6-5(a). The same thing might occur if the price of a substitute product, such as skirts or corduroy pants, were to increase, because that would put jeans at a competitive advantage. If two goods are substitutes, a rise in the price of one of them will tend to cause the demand curve for the other one to shift outward (to the right). Similarly, if a product that is complementary to jeans (perhaps a certain type of belt)

FIGURE 6-5 SHIFTS IN A DEMAND CURVE



A rise in consumer income, an increase in advertising, or a rise in the price of a competing product can all produce a rightward (outward) shift of the demand curve for a product, as depicted by the shift from the black curve, D_0D_0, to the orange curve, D_1D_1 in Panel (a). This means that at any fixed price (say, \$25), the quantity of the product demanded rises. (In the figure, it rises from 40,000 to 60,000 units.) Similarly, a fall in any of the variables, such as consumer income, will produce a leftward (inward) shift in the demand curve, as in Panel (b).

becomes more expensive, we would expect the demand curve for jeans to shift to the left, as in Figure 6-5(b). In summary:

A demand curve is expected to shift to the right (outward) if consumer incomes rise, if tastes change in favor of the product, if substitute goods become more expensive, or if complementary goods become cheaper. A demand curve is expected to shift to the left (inward) if any of these factors goes in the opposite direction.

THE TIME DIMENSION OF THE DEMAND CURVE AND ECONOMIC DECISION MAKING

One more important feature of a demand curve does not show up on a graph. A demand curve indicates, at each possible price, the quantity of the good that is demanded *during a particular period of time*. That is, all of the alternative prices considered in a demand curve must refer to *the same* time period. Economists do not compare a price of $10 for January with a price of $8 for September.

This feature imparts a peculiar character to the demand curve and complicates calculations of statistical estimates. Actual, observed data show different prices and quantities only for different dates. Why, then, do economists adopt this apparently peculiar approach? The answer is that the time dimension of the demand curve is dictated inescapably by the logic of decision making.

When a business undertakes to find the best price for one of its products for, say, the following 6 months, it must consider the range of alternative prices available to it for that 6-month period and the consequences of each possible choice. For example, if management is reasonably certain that the best price lies somewhere between $3.50 and $5.00, it should perhaps consider each of four possibilities, $3.50, $4.00, $4.50, and $5.00, and estimate how much it can expect to sell at each of these potential prices during the 6-month period. The result of these estimates may appear in a format similar to that of the table below.

Potential Price	Expected Quantity Demanded
$3.50	75,000
4.00	73,000
4.50	70,000
5.00	60,000

This table supplies managers with information that they need to make an optimal pricing decision. It also contains precisely the information an economist uses to draw a demand curve.

The demand curve describes a set of hypothetical quantity responses to a set of potential prices, only one of which can actually be charged. All of the points on the demand curve refer to alternative possibilities for the *same* period of time—the period for which the decision is to be made.

Thus, a demand curve of the sort just described is not just an abstract notion that is useful primarily in academic discussion. Rather, it offers precisely the information that businesses need for rational decision making. However, as already noted, the fact that all points on the demand curve are hypothetical possibilities for the same period of time causes problems for statistical estimation of demand curves. These problems are discussed in the appendix to this chapter.

TWO ILLUSTRATIVE APPLICATIONS OF ELASTICITY ANALYSIS

This chapter started out with two illustrative cases. We are now in a position to show how the elasticity concept helps to resolve them.

EXAMPLE I: THE SALES-TAX REDUCTION

Recall that the first example involved a 20 percent reduction in the excise taxes on cigarettes and beer in the state of Confusion. The politicians who arranged the tax cut were pleasantly surprised to find that tax revenues fell by less than 20 percent. They also found that the decline in revenue from the beer tax was less than that for cigarettes, which puzzled them.

You should by now easily see the explanations. Officials in Confusion had failed to take account of the elasticities of demand for the two products. Both products clearly had some price elasticity of demand. The fall in retail prices that followed from the tax cut stimulated additional purchases and therefore produced additional tax receipts, which partially offset the direct revenue loss attributable to the tax cut.

The retail prices of the two goods fell by similar amounts after the tax cut, but beer sales increased more than cigarette sales. This suggests that the demand curve for beer must have been more elastic than the demand curve for cigarettes. This is plausible because consumers can probably find closer substitutes for beer (say, wine coolers) than for cigarettes. As a consequence, beer-tax revenues fell by less than cigarette-tax revenues.

EXAMPLE II: *POLAROID V. KODAK*

In 1989, a lengthy trial determined how much money Kodak owed Polaroid for its patent infringement when it sold instant cameras and film for a 10-year period from 1976 to 1986. The key issue was how much profit Polaroid had lost as a result of Kodak's entry into this field. The concepts of price elasticity of demand and cross elasticity of demand both played crucial roles.

Estimates of the price elasticity of demand were important in determining whether the explosive growth in instant camera sales from 1976 to 1979 was mainly attributable to the fall in price that resulted from Kodak's competition or to Kodak's reputation and its access to additional retail outlets. In the latter case, Polaroid might actually have benefited from Kodak's entry, rather than lost profits as a result.

After 1980, sales of instant cameras and film began to drop sharply. On this issue, the *cross elasticity* of demand between instant and noninstant cameras and film was crucial to the explanation. Why? Because the decline in the instant camera market occurred just as the prices of 35-millimeter cameras, film, developing, and printing all began to fall significantly. If the decrease in cost of 35-millimeter photography caused the decline in Polaroid's overall sales, then it was not the fault of Kodak's instant photography activity. Consequently, the amount that Kodak would be required to pay to Polaroid would be significantly smaller. On the other hand, if the cross elasticity of demand between conventional photography prices and the demand for instant cameras and film was low, then the cause of the decline in Polaroid's sales might have been Kodak's patent-infringing activity—adding to the damage payments to which Polaroid was entitled.

Yet a third elasticity issue was central to the trial. How much would Polaroid's total revenue have increased as a result of the film-price increase it claimed it would have adopted in the absence of illegal competition from Kodak? As you

now know, the crucial matter here was the price elasticity of demand for instant camera film. Polaroid tried to show that this elasticity figure was low to support its contention that the rise in price would have raised its revenues a great deal—meaning that Kodak had damaged it quite severely. Kodak, on the other hand, offered statistical evidence suggesting that the elasticity of demand for instant film was high to support its own claim that a higher price would have brought in little, if any, additional revenue.

On the basis of its calculation of all of the relevant elasticities, Polaroid at one point claimed that Kodak was obligated to pay it $9 billion or more. Kodak claimed that it owed Polaroid only (!) something in the neighborhood of $450 million. Much was obviously at stake. The judge's verdict came out with a number very close to Kodak's figure.

SUMMARY

1. To measure the responsiveness of the quantity demanded to price, economists calculate the **elasticity of demand,** which is defined as the percentage change in quantity demanded divided by the percentage change in price.
2. If demand is *elastic* (elasticity exceeds 1), a rise in price will reduce total expenditures. If demand is *unit elastic* (elasticity equal to 1), a rise in price will not change total expenditures. If demand is *inelastic* (elasticity less than 1), a rise in price will increase total expenditure.
3. Demand is not a fixed number. Rather, it is a relationship showing how quantity demanded is affected by price and other pertinent influences. If one or more of these other variables change, the demand curve will shift.
4. Goods that make each other more desirable (hot dogs and mustard, wristwatches and watch straps) are called **complements.** When consumers want less of one good as they get more of another (steaks and hamburgers, Coke and Pepsi), economists call those goods **substitutes.**
5. **Cross elasticity of demand** is defined as the percentage change in the quantity demanded of one good divided by the percentage change in the price of another good. Two substitute products normally have a positive cross elasticity of demand. Two complementary products normally have a negative cross elasticity of demand.
6. A rise in the price of one of two substitute products can be expected to *shift the demand curve* of the other product to the right. A rise in the price of one of two complementary goods is apt to shift the other's demand curve to the left.
7. All points on a demand curve refer to the *same* time period—the time during which the price will be in effect.

KEY TERMS

(Price) elasticity of demand
Elastic, inelastic, and unit-elastic demand curves
Complements
Substitutes
Cross elasticity of demand
Shift in a demand curve

QUESTIONS FOR REVIEW

1. What variables besides price and advertising are likely to affect the quantity demanded of a product?
2. Describe the probable shifts in the demand curves for:
   a. airplane trips when the airlines' on-time performance improves.
   b. automobiles when airplane fares rise.
   c. automobiles when gasoline prices rise.
   d. electricity when the average temperature in the United States falls during a particular year. (Note: The demand curve for electricity in Maine and the demand curve for electricity in Florida should respond in different ways. Why?)

3. Taxes on particular goods discourage their consumption. Economists therefore say that such taxes "distort consumer demands." In terms of the elasticity of demand for the commodities in question, what sort of goods would you choose to tax to achieve the following objectives?
   a. Collect a large amount of tax revenue.
   b. Distort demand as little as possible.
   c. Discourage consumption of harmful commodities.
   d. Discourage production of polluting commodities.
4. Explain why elasticity of demand is measured in *percentages*.
5. Explain why the elasticity of demand formula normally eliminates minus signs.
6. Give examples of commodities whose demand you expect to be elastic and some whose demand you expect to be inelastic.
7. Explain why the elasticity of a straight-line demand curve varies from one part of the curve to another.
8. A rise in the price of a certain commodity from $15 to $20 reduces quantity demanded from 20,000 to 5,000 units. Calculate the price elasticity of demand.
9. If the price elasticity of demand for gasoline is 0.3, and the current price is $1.20 a gallon, what rise in the price of gasoline will reduce its consumption by 10 percent?
10. A rise in the price of a product whose demand is elastic will reduce the total revenue of the firm. Explain.
11. Name some events that will cause a demand curve to shift.
12. Which of the following product pairs would you expect to be substitutes and which would you expect to be complements?
    a. Shoes and sneakers.
    b. Gasoline and big cars.
    c. Bread and butter.
    d. Instant camera film and regular camera film.
13. For each of the previous product pairs, what would you guess about the products' cross elasticity of demand?
    a. Do you expect it to be positive or negative?
    b. Do you expect it to be a large or small number? Why?
14. Explain why the following statement is true: "A firm with a demand curve that is inelastic at its current output level can always increase its profits by raising its price and selling less." (*Hint*: Refer back to the discussion of elasticity and total expenditure on pages 133–135.)

APPENDIX

STATISTICAL ANALYSIS OF DEMAND RELATIONSHIPS

The peculiar time dimension of the demand curve, in conjunction with the fact that many variables other than price can influence quantity demanded, makes it surprisingly hard to derive a product's demand curve from statistical data. It can be done, but the task is full of booby traps and can usually be carried out successfully only by using advanced statistical methods. Let us see why these two characteristics of demand curves cause problems.

The most obvious way to go about estimating a demand curve statistically is to collect a set of figures on prices and quantities sold in different periods, like those given in Table 6-2. These points can then be plotted on a diagram with price and quantity on the axes, as shown in Figure 6-6. One can then draw in a line (the dotted line *TT*) that connects these points reasonably well. This line may appear to approximate the demand curve, but

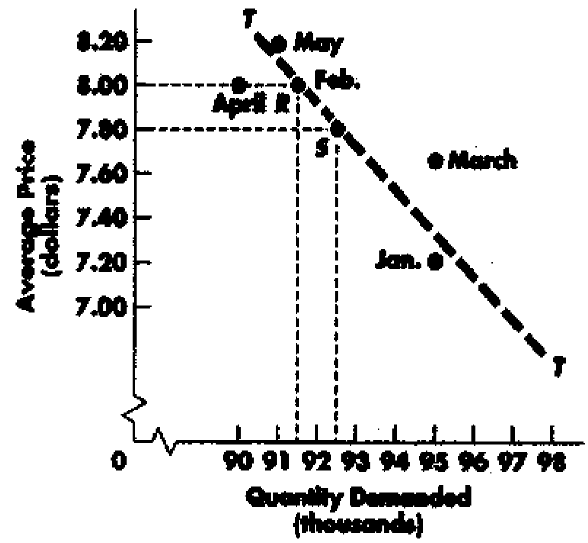
TABLE 6-2

HISTORICAL DATA ON PRICE AND QUANTITY

	January	February	March	April	May
Quantity sold	95,000	91,500	95,000	90,000	91,000
Price	$7.20	$8.00	$7.70	$8.00	$8.20

FIGURE 6-6 PLOT OF HISTORICAL DATA ON PRICE AND QUANTITY

The dots labeled Jan., Feb., and so on represent actual prices and quantities sold in the months indicated. The orange line, *TT*, is drawn to approximate the dots as closely as possible.



unfortunately, line *TT*, which summarizes the historical data, may bear no relationship to the true demand curve.

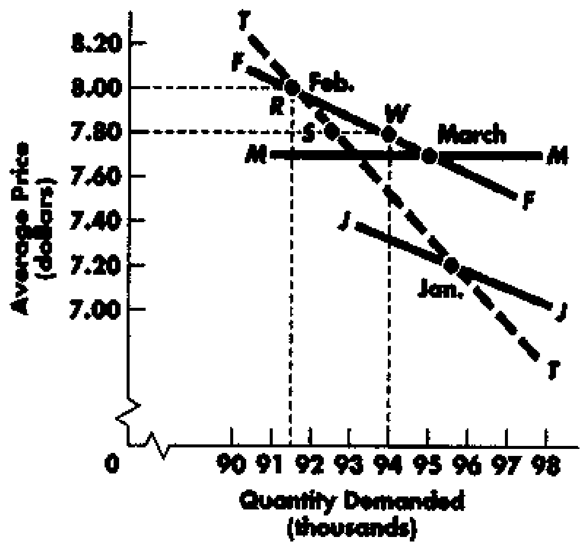
You may notice at once that the prices and quantities represented by the historical points in Figure 6-6 refer to different periods of time and that they all have been *actual*, not *hypothetical*, prices and quantities at *the same* time. The distinction is significant. Over the period covered by the historical data, the true demand curve, which is what an economist really wants, may well have shifted because of changes in some of the other variables affecting quantity demanded.

What actually happened may be as shown in Figure 6-7. In January, the demand curve was given by *JJ*, but by February the curve had shifted to *FF*, by March to *MM*, and so on. This figure shows a separate and distinct demand curve for each of the relevant months, and none of them need have any resemblance to the plot of historical data, *TT*.

In fact, the slope of the historical plot curve, *TT*, can be very different from the slopes of the true

FIGURE 6-7 PLOT OF HISTORICAL DATA AND TRUE DEMAND CURVES FOR JANUARY, FEBRUARY, AND MARCH

An analytical demand curve shows how quantity demanded in a particular month is affected by the different prices considered during that month. In the case shown, the true demand curves are much flatter (more elastic) than the line plotting historical data. This means that a cut in price will induce a far greater increase in quantity demanded than the historical data suggest.



underlying demand curves, as is the case in Figure 6-7. This means that the decision maker can be seriously misled if she selects her price on the basis of the historical data. She may, for example, think that demand is quite insensitive to changes in price (as line *TT* seems to indicate), and so she may reject the possibility of a price reduction; in fact, the true demand curves show that a price reduction would increase quantity demanded substantially.

For example, if she were to charge a price of $7.80 rather than $8.00 in February, the historical plot would lead her to expect a rise in quantity demanded of only 1,000 units. (Compare point *R*, with sales of 91,500 units, and point *S*, with sales of 92,500 units, in Figure 6-6.) However, the true demand curve for February (line *FF* in Figure 6-7) indicates an increase in sales of 2,500 units (from point *R*, with sales of 91,500 units, to point *W*, with sales of 94,000 units). A manager who based her decision on the historical plot, rather than on the true demand curve, might be led into serious error. Nevertheless, it is astonishing how often people make this mistake in practice, even when using apparently sophisticated techniques.

AN ILLUSTRATION: DID THE ADVERTISING PROGRAM WORK?

Some years ago one of the nation's largest producers of packaged foods conducted a statistical study to judge the effectiveness of its advertising expenditures, which amounted to nearly $100 million a year. A company statistician collected year-by-year figures on company sales and advertising outlays and discovered, to his delight, that they showed a remarkably close relationship to one another: quantity demanded rose as advertising rose. The trouble was that the relationship seemed just too perfect. In economics, data about demand and any one of the elements that influence it almost never make such a neat pattern. Human tastes and other pertinent influences are just too variable to permit such regularity.

Suspicious company executives asked one of the authors of this book to examine the analysis. A little thought showed that the suspiciously close statistical relationship between sales and advertising expenditures resulted from a disregard for the principles just presented. The investigator had in fact constructed a graph of *historical* data on sales and advertising expenditure, analogous to *TT* in Figures 6-6 and 6-7 and therefore not necessarily similar to the truly relevant relationship.

The stability of the relationship actually arose from the fact that, in the past, the company had based its advertising outlays on its sales, automatically allocating a fixed percentage of its sales revenues to advertising. The *historical* relationship between advertising and demand therefore described only the company's budgeting practices, not the effectiveness of its advertising program. If the firm's management had used this curve in planning future advertising campaigns, it might have made some regrettable decisions. *Moral:* Avoid the use of historical curves like *TT* in making economic decisions.

CHAPTER 7

PRODUCTION, INPUTS, AND COST: BUILDING BLOCKS FOR SUPPLY ANALYSIS

You realize this cost figure is only an estimate. Of course, the actual cost will be higher.

Architect to a client in a *New Yorker* cartoon

The prices and amounts of goods sold in markets do not depend on demand alone. They also depend on *supply,* or the quantities of those goods that firms offer for sale in the market. The supply (output) decisions of firms depend, in turn, on their costs. Thus, to understand the way markets work, we must now turn to the determinants of the costs of production.

This chapter does precisely that. A firm's costs depend on the quantities of labor, raw materials, machinery, and other inputs that it buys—and on the prices of each. But that observation raises a question: How does the firm decide how much of each of the various *inputs* to buy? This chapter examines how a business can select the optimal combination of inputs, that is, the combination that enables it to produce whatever output it decides upon at minimum cost and, hence, with maximum profitability.

To make the analysis easier to follow, we approach this task in two stages. The first part of the chapter deals with the simple case in which the firm can vary the quantity of only a single input. This unrealistic assumption vastly simplifies the analysis and will enable us to understand more easily the three key issues of this chapter: how the quantity of input affects the quantity of output, how the firm selects the optimal quantity of an input, and how these input decisions give the firm the cost information it needs to decide on output and price.

The second part of the chapter goes over the same territory for the more realistic case in which the firm is free to choose the quantities of several inputs at once. Many new insights emerge from the multi-input analysis, but they will be easier to absorb after you have mastered single-input analysis.[1]

[1]NOTE: Some instructors may prefer to postpone this part until later in the course.

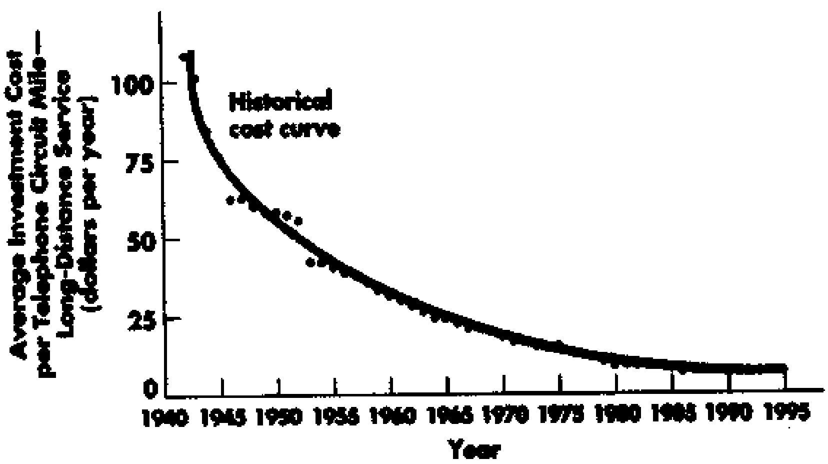
APPLICATION: ARE LARGER FIRMS MORE EFFICIENT?

Economies of large-scale production are thought to be a pervasive feature of modern industrial society. Automation, assembly lines, and sophisticated machinery often reduce production costs dramatically. But if equipment of such enormous capacity requires a very large investment, small companies will be unable to reap many of these benefits of modern technology. Only large firms will be able to take advantage of the associated savings in costs. Where firms can take advantage of such *economies of scale*, as economists call them, production costs per unit will decline as output expands.

But this favorable relationship between large size and low costs does not characterize every industry. Sometimes a court of law must decide whether a giant firm should be broken up into smaller units. The most celebrated case of this kind involved American Telephone and Telegraph Co., which had a monopoly on the nation's phone service for nearly 50 years. In 1982, AT&T settled a government antitrust suit by agreeing to divest 22 operating units known today as the *Baby Bells*. Those who had urged a breakup of AT&T argued that such a giant firm concentrated economic power and deprived consumers of the benefits of competition. But those who opposed the breakup pointed out that, if AT&T could create significant economies of scale, smaller firms would be much less efficient producers than a larger one. Who was right? In order to settle the issue, one would have to know whether significant economies of scale affected AT&T's business before the divestiture.

Sometimes data like those shown in Figure 7-1 are offered to the courts when they consider such cases. These figures, provided by AT&T, indicate that as the volume of telephone messages rose after 1942, the capital cost of long-distance communication by telephone dropped enormously. Economists maintain that, while this graph may be valid evidence of efficiency and innovation in the telecommunications industry, it does *not* constitute legitimate evidence, one way or another, about the presence of economies of scale. At the end of this chapter, we will study precisely what is wrong with such evidence and what sort of evidence really is required to determine whether a very large firm can indeed produce more efficiently.

FIGURE 7-1 HISTORICAL COSTS FOR LONG-DISTANCE TELEPHONE TRANSMISSIONS



By 1995, the dollar cost per circuit mile of long-distance calls had fallen below 8 percent of its level in 1942. Because prices had more than tripled in that period, the decline in *real* cost was even more sensational. Yet this diagram of historical costs is not legitimate evidence *one way or the other* about economies of scale in telecommunications.

SOURCE: AT&T. Recent figures are estimates.

PRODUCTION, INPUT CHOICE, AND COST WITH ONE VARIABLE INPUT

TABLE 7-1

FLO'S TOTAL PHYSICAL PRODUCT SCHEDULE

Corn Input (40-lb. bags)	Chicken Output (lbs.)
0	0.0
1	14.0
2	36.0
3	66.0
4	**100.0**
5	130.0
6	156.0
7	175.0
8	184.0
9	185.4
10	180.0
11	165.0
12	144.0

Note: Data of the sort provided in this table do not represent the farmers' subjective opinion. They are objective information of the sort that a poultry scientist can and does supply from experimental evidence.

While all real businesses use many different inputs, we will begin our discussion with the unrealistic case of only a single variable input. In doing so, we are trying to replicate in our theoretical analysis what physicists or biologists do in the laboratory when they conduct a *controlled* experiment—changing just one variable at a time so as to study the influence of that variable in isolation.

TOTAL, AVERAGE, AND MARGINAL PHYSICAL PRODUCTS

We begin the analysis with the first of the chapter's three topics: the relationship between the quantity of production and the quantity of inputs utilized. Consider, as an example, the case of Florence Farmer, who operates a small poultry business, and the amount of corn that she feeds her chickens. Ultimately, she can vary all the input quantities. She can buy more baby chicks, feed them different grains, give the chickens growth hormones, buy more land and chicken coops, and so on. But suppose for the moment that Flo's only choice is how much corn to feed the flock.

Flo has studied the relationship between the quantity of corn input and the output of poultry meat. In doing so, she has learned how much additional output she gets if she uses various amounts of corn. The relevant data are displayed in Table 7-1.

The table begins by confirming the common-sense observation that chickens cannot be raised without food. Thus, output is zero when Flo provides zero corn input (see the first line of the table). After that, the table shows the increase in chicken output that additional corn yields. For instance, with an input of four 40-pound bags of corn per week, output is 100 pounds of chicken. Eventually, however, a saturation point is reached beyond which additional corn overfeeds the chickens and causes some to become ill. After nine bags of corn per week, more corn actually reduces the chicken output.

The firm's **TOTAL PHYSICAL PRODUCT (TPP)** is the amount of output it obtains in total from a given quantity of input.

The data in Table 7-1 are portrayed graphically in Figure 7-2, which is called a **total physical product (TPP)** curve. This curve shows how much chicken Florence Farmer can produce with different quantities of corn, holding the quantities of all other inputs constant.

The **AVERAGE PHYSICAL PRODUCT (APP)** is the total physical product (TPP) divided by the quantity of input. Thus, APP = TPP/X where X = the quantity of input.

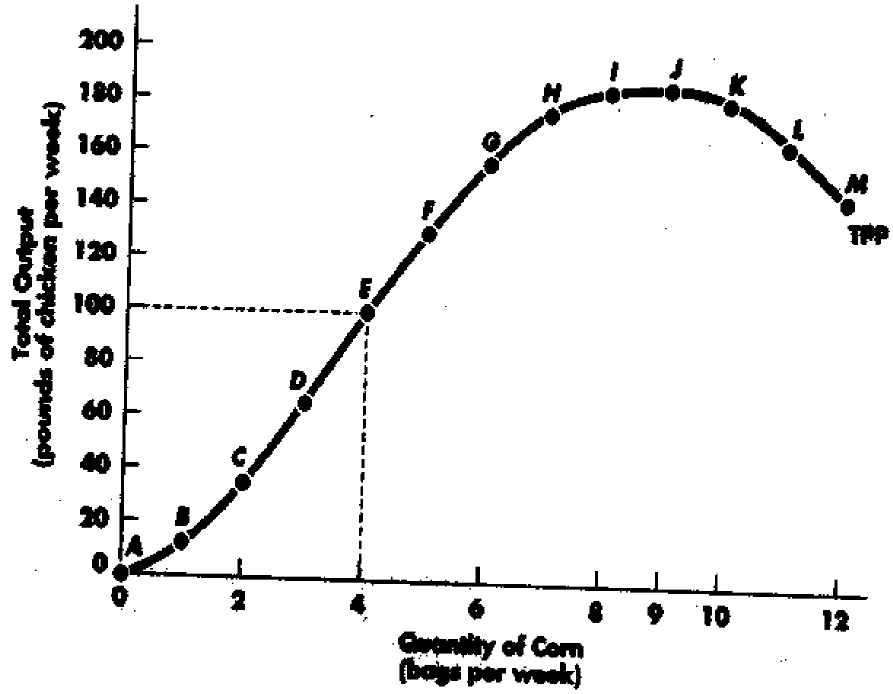
Table 7-2 adds two other physical product concepts. **Average physical product (APP)** measures output per unit of input; it is simply total physical product divided by the quantity of variable input used. On Flo's farm, it is total chicken output divided by the number of bags of corn. APP is shown in the next to last column of Table 7-2. For example, since four bags of corn yield 100 pounds of chicken, the APP of four bags of corn is 100/4, or 25 pounds per bag of corn.

The **MARGINAL PHYSICAL PRODUCT (MPP)** of an input is the increase in total output that results from a 1-unit increase in the input, holding the amounts of all other inputs constant.

To decide how much corn to feed her flock, Florence must know how much *additional* chicken output to expect from each *additional* bag of corn. This concept is known as **marginal physical product (MPP)**, and she can calculate it from the data on total physical product using the same method we introduced to derive marginal utility from total utility in Chapter 5. For example, the marginal physical product of the fourth bag of corn is the total output of chicken when Flo

FIGURE 7-2 TOTAL PHYSICAL PRODUCT WITH DIFFERENT QUANTITIES OF CORN

This graph shows how Flo's chicken output varies as she uses more and more corn. (Other inputs, such as labor, are held constant in this graph.)



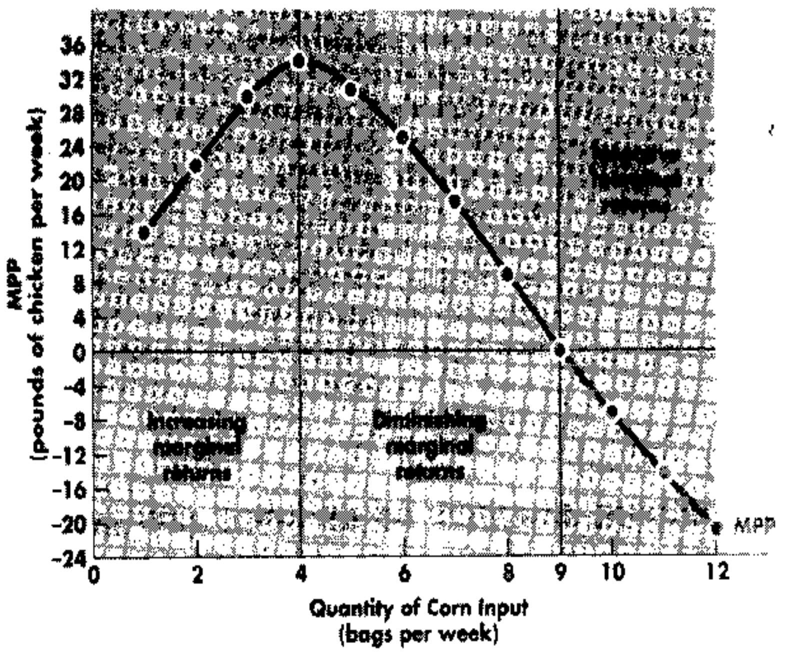
uses four bags of corn *minus* the total output when she uses three bags (100 − 66 = 34 pounds). The other MPP entries in the third column of Table 7-2 are calculated in exactly the same way. Figure 7-3 displays these numbers in a graph called a *marginal physical product curve*.

TABLE 7-2 FLO'S SCHEDULES FOR TOTAL PHYSICAL PRODUCT, MARGINAL PHYSICAL PRODUCT, AVERAGE PHYSICAL PRODUCT, AND MARGINAL REVENUE PRODUCT OF CORN

Corn Input (bags)	Total Physical Product (chicken output in lbs.)	Marginal Physical Product (lbs. per bag)	Average Physical Product (lbs. per bag)	Marginal Revenue Product (dollars)
0	0.0		—	
1	14.0	14.0	14.0	10.50
2	36.0	22.0	18.0	16.50
3	66.0	30.0	22.0	22.50
4	100.0	34.0	25.0	25.50
5	130.0	30.0	26.0	22.50
6	156.0	26.0	26.0	19.50
7	175.0	19.0	25.0	14.25
8	184.0	9.0	23.0	6.75
9	185.4	1.4	20.6	1.05
10	180.0	−5.4	18.0	−4.05
11	165.0	−15.0	15.0	−11.25
12	144.0	−21.0	12.0	−15.75

FIGURE 7-3 FLO'S MARGINAL PHYSICAL PRODUCT (MPP) CURVE

This graph of marginal physical product (MPP) shows how much *additional* chicken Flo gets from each additional bag of corn.



THE "LAW" OF DIMINISHING MARGINAL RETURNS

The shape of the marginal physical product curve in Figure 7-3 is significant for this analysis. Until input reaches four bags of corn, the marginal physical product of corn is *increasing;* between four bags and nine bags it is *decreasing*, but still *positive;* beyond nine bags, the MPP of corn actually becomes *negative*. The graph is divided into three zones to illustrate these three cases.

Note that the marginal returns to corn increase at first and then diminish. This is the typical pattern.

In the left zone, the region of increasing marginal returns, each additional bag of corn adds more to TPP than the previous bag did. This corresponds to the curve between points *A* and *E* in Figure 7-2, where the TPP curve rises increasingly rapidly. In the middle zone, the region of diminishing (but still positive) marginal returns, each additional bag of corn adds less to TPP than the previous bag added. This corresponds to the curve between points *E* through *I* in Figure 7-2, where the TPP curve continues to rise, but at a diminishing rate. Finally, in the right zone of negative marginal returns (input quantities greater than nine bags), additional corn actually reduces production by making the chickens ill.

The "law" of diminishing marginal returns, which has played a key role in economics for two centuries, states that an increase in the amount of any one input, *holding the amounts of all others constant*, ultimately leads to lower marginal returns to the expanding input.[2] The so-called *law* is simply based on some observed facts; it is not a theorem deduced analytically.

[2]The "law" is generally credited to Anne Robert Jacques Turgot (1727–1781), one of the great Comptrollers-General of France before the revolution, whose liberal policies, it is said, represented the old regime's last chance to save itself. With characteristic foresight, the king fired Turgot.

The reason why returns to a single input usually diminish is a result of what can be referred to as the "law" of variable input proportions. When the quantity of one input increases while all others remain constant, the input whose quantity increases gradually becomes more and more abundant compared with the others. (For example, the proportion of corn to other foods increases.) As Florence Farmer uses more and more corn with fixed quantities of other foods, the chickens' diet becomes unbalanced so that adding yet more corn does little good and eventually harms them. At this point, the marginal physical product of corn becomes *negative*. Plenty of real-world cases seem to display the law of variable input proportions. In China, for instance, farmers have been using more and more fertilizer in an effort to produce larger grain harvests to feed the country's burgeoning population. While that country's consumption of fertilizer is four times higher than it was 15 years ago, its grain output has increased by only 50 percent. This certainly suggests that fertilizer use has reached the zone of diminishing returns.

THE OPTIMAL QUANTITY OF AN INPUT AND DIMINISHING RETURNS

We now have all the tools we need to see how a firm can decide on the quantity of input that will maximize its profits. For this purpose, look again at the first and third columns of Table 7-2, which show the farm's marginal physical product schedule. Suppose that corn costs $10 per bag, and that Flo can sell her chicken for $0.75 per pound.

Now suppose that Flo is considering using just one bag of corn. Is this optimal, that is, does it maximize profits? The answer is no, because the marginal physical product of a second bag of corn is 22 pounds of chicken (the second entry in the marginal physical product column of Table 7-2). At a price of $0.75 per pound, this extra output would add $16.50 to revenue. Since the added revenue exceeds the $10 cost of the second bag of corn, the farm comes out ahead by $16.50 minus $10.00, or $6.50.

The **MARGINAL REVENUE PRODUCT (MRP)** of an input is the additional revenue that the producer earns from the increased sales when it uses an additional unit of the input.

The additional *money revenue* that a firm receives when it increases the quantity of some input by 1 unit is called the input's **marginal revenue product.** If Flo's chicken sells at a fixed price of $0.75 per pound, the marginal revenue product (MRP) of the input equals its marginal physical product (MPP) multiplied by the price:

$$\text{MRP} = \text{MPP} \times \text{Price of output}$$

For example, we have just shown that the marginal revenue product of the second bag of corn is $16.50, which we obtained by multiplying the MPP of 22 pounds by the price of $0.75 per pound. The other (MRP) entries in the last column of Table 7-2 are calculated in the same way. The concept of MRP enables us to formulate a simple rule for the optimal use of any input. Specifically:

When the marginal revenue product of an input exceeds its price it pays the firm to use more of that input. Similarly, when the marginal revenue product of the input is less than its price it pays the firm to use less.

Let us test this rule in the case of Florence's farm. We have observed that two bags of corn cannot be enough because the MRP of a second bag ($16.50) exceeds its price ($10). What about a third bag? Table 7-2 shows that the MRP of the third bag ($0.75 × 30 = $22.50) also exceeds its price; thus, stopping at three bags also is not optimal. This remains true through the seventh bag of corn, whose MRP

of $14.25 still exceeds its $10 price. The same cannot be said of an eighth bag, however. The eighth bag is not a good idea because its MRP, only $6.75, is less than its $10 cost. Thus, the optimal quantity of corn for Flo to purchase each week is seven bags, yielding a total output of 175 pounds of chicken.

Notice the crucial role of diminishing returns in this analysis. When the marginal *physical* product of corn begins to decline, the money value of that product falls, as well; that is, the marginal *revenue* product also declines. The producer always profits by expanding input use until diminishing returns set in and reduce the MRP to the price of the input. So Flo should stop increasing her corn purchases when MRP falls to the price of corn.

A common expression suggests that it does not pay to continue doing something "beyond the point of diminishing returns." As we see from this analysis, quite to the contrary, it normally *does* pay to do so! The firm has employed the proper amount of input only when diminishing returns reduce the marginal revenue product of the input to the level of its price, because then the firm will be wasting no opportunity to *add* to its total profit. Thus, the optimal quantity of an input is that at which MRP equals its price (P). In symbols,

$$\text{MRP} = P \text{ of input}$$

Notice that the logic of this analysis is exactly the same as in our discussion of marginal utility and price in Chapter 5. Florence Farmer is trying to maximize profits—the difference between the *total* revenue her corn input yields and the *total* cost of buying that input. To do so, she must increase the usage of corn up to the point where its price equals its marginal revenue product, just as an optimizing consumer kept buying until price equaled marginal utility.

COST CURVES AND INPUT QUANTITIES

We turn now to the third of the three main topics of this chapter: how the firm's cost relationships are derived from the input decisions that we have just explained. The cost relationships we will discuss are the ones that we will need to analyze the firm's output and pricing decisions, the subject of the next chapter. There we will study the last of the main components of our analysis of the market mechanism: how much of its commodity the profit-maximizing firm should produce.

The quantity of output that is most desirable for the firm clearly depends on the way in which costs change when output varies. Such information is typically displayed in the form of *cost curves*. Indeed, since we will use marginal analysis again in our discussion, we will need three different cost curves: *the total cost curve, the average cost curve, and the marginal cost curve.*

These curves follow directly from the nature of production. Production technology dictates the amount of corn that Flo uses to produce a given quantity of chicken. Based on corn usage and the price of corn, Flo can determine how much it will cost to produce some given amount of chicken. Therefore, the needed cost relationships depend directly on the production relationships we have just discussed. The calculation of the firm's total costs from its physical product schedule assumes that the price of corn is beyond the control of the firm and that the quantities of all inputs other than corn are fixed.

The method is simple: For each quantity of output, record from Table 7-1 or Figure 7-2 the amount of corn required to produce it. Then multiply that quantity of corn by its price.

In addition to the cost of corn, Flo must spend money on other inputs. Furthermore, her costs must include the *opportunity costs* of any inputs that Flo herself contributes (such as soy beans that she grew herself) because she could have sold them to other farmers instead of feeding them to her chickens. Since we are simplifying matters by focusing exclusively on corn input for now, however, let us assume for the next several pages that all other inputs are free. Later, we will bring in the costs of other inputs.

Since corn costs $10 per bag, the data in Table 7-1 (page 149) lead directly to the total costs shown in Table 7-3. For example, the fourth row of Table 7-1 indicates that Flo needs three bags of corn to produce 66 pounds of chicken. At $10 per bag, this input costs $30, which is the third entry in the last column of Table 7-3. The other numbers in Table 7-3 are derived similarly. In general:

The total product curve specifies the input quantities needed to produce any given output. From those input quantities and the prices of the inputs, we can determine the *total cost* (TC) of producing any level of output. (This amount must include any opportunity costs that arise in that production activity.) Thus, the relationship of total cost to output is determined by the technological production relationships between inputs and outputs and by input prices.

Later analysis will require two other cost curves—average cost and marginal cost. Flo can calculate them directly from the total cost curve just as Table 7-2 calculated average and marginal physical product from total physical product. For any given output, *average cost* is defined as total cost divided by quantity produced. For example, Table 7-3 shows that the total cost of producing 100 pounds of chicken is $40, so that the average cost is $40/100, or $0.40.

Similarly, the *marginal cost* is defined as the increase in total cost that results from the production of an additional pound of chicken. For example, the marginal cost of the 100th pound of chicken is the difference between the total cost of producing 100 pounds ($40) and the total cost of producing 99 pounds (say, $39.70, although the table doesn't show this number); this difference equals $0.30. Figure 7-4 shows all three curves, the total, average, and marginal cost curves. The TC curve is generally assumed to rise fairly steadily as the firm's output

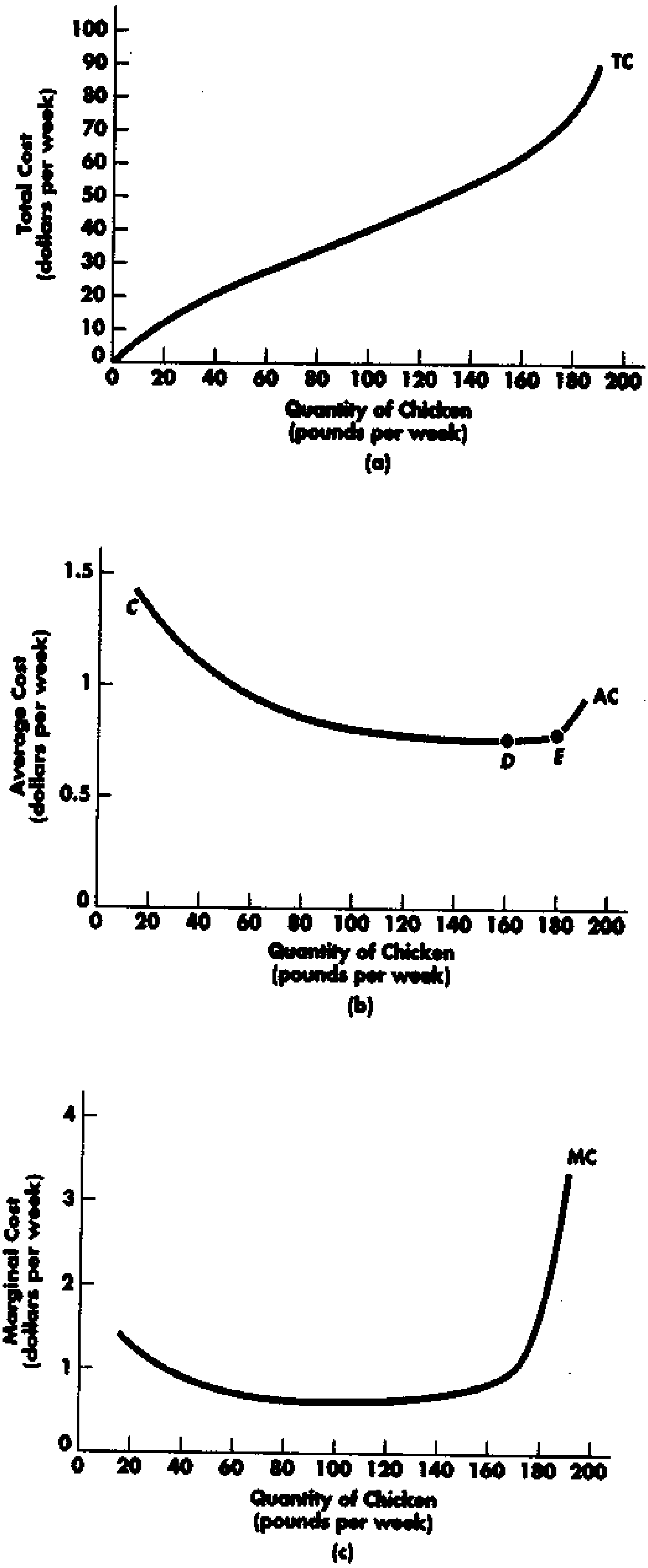
TABLE 7-3 A PORTION OF FLO'S TOTAL COST SCHEDULE[a]

Chicken Output (lbs. per week)	Corn Input (bags per week)	Total Cost (cost of corn, $ per week)
0	0	$ 0
14	1	10
36	2	20
66	3	30
100	4	40
130	5	50
156	6	60
175	7	70
184	8	80

[a]Obtained from the production data in Table 7-1, assuming that corn is the only variable input.

FIGURE 7-4 TOTAL, AVERAGE, AND MARGINAL COSTS

Flo's costs are derived from the total cost schedule in Table 7-3.



increases. After all, one cannot expect to produce 100 pounds of chicken at a lower total cost than 80 pounds. The AC curve and the MC curve are both shaped roughly like the letter *U*—first going downhill, then gradually turning uphill again. We will explore the reason for the *U*-shape of the curves later in the chapter.

So far, we have ignored the money that Florence Farmer must spend on inputs other than corn. Of course, these other inputs will not be available free, so their costs should be treated as constants, not as zero, in the present analysis. Since such fixed costs are quite important in reality, we will examine them more closely.

FIXED COSTS AND VARIABLE COSTS

A **FIXED COST** is the cost of an input whose quantity does not rise when output goes up, one that the firm requires to produce any output at all. The total cost of such indivisible inputs does not change when the output changes. Any other cost of the firm's operation is called a **VARIABLE COST.**

Total, average, and marginal costs are often divided into two components: fixed costs and variable costs. A **fixed cost** is the cost of the smallest (least expensive) batch of inputs that the firm can buy in order to produce any output at all. These costs are called *fixed* because the total amount of money spent on the inputs does not increase when output rises, at least up to some point. The total cost of such inputs remains the same whether output is 5 units or 500 units.

Such fixed costs arise because a firm cannot reduce the quantities of some inputs, even by producing very little output. For example, it takes at least one automobile to run a taxi business, and that input is the same whether the cab driver serves 1 rider a day or 50. Obviously, the total cost of such an input does not change as the firm varies its output, at least within some range. (Beyond some point, if the number of riders grows sufficiently large, the taxi business will need a second automobile.) Costs that are not fixed in this sense, that is, those that increase when the firm's output rises, are called **variable costs.**

To illustrate the difference between fixed costs and variable costs, compare the cost of a railroad's fuel with that of building the track. To operate between St. Louis and Kansas City, a railroad must lay a set of tracks. It cannot lay half or a quarter of a set of tracks. We therefore call such an input *indivisible.*

At any output *greater than zero* the railroad has no choice but to pay the fixed cost of its track. Because the track is indivisible, its construction cost will be the same whether one train per month or five trains per day travel the route.[3] Thus, up to a point, track construction cost is unaffected by output size, that is, by the volume of traffic. On the other hand, as more trains pass over those tracks, the railroad must pay a higher fuel bill. Economists say that its fuel costs are variable, but its track-building cost is fixed.

Although variable costs are only part of total costs (fixed plus variable costs), variable cost curves have the same shape as the total cost curve shown in Figure 7-4. In contrast, the curves of *total fixed costs* (TFC) and *average fixed costs* (AFC) have very special shapes, illustrated in Figure 7-5. Naturally, TFC remains the same whether the firm produces a little or a lot—so long as it produces anything at all.[4] As a result, any TFC curve is horizontal like the one in Figure 7-5(a). It has the same height at every output.

---

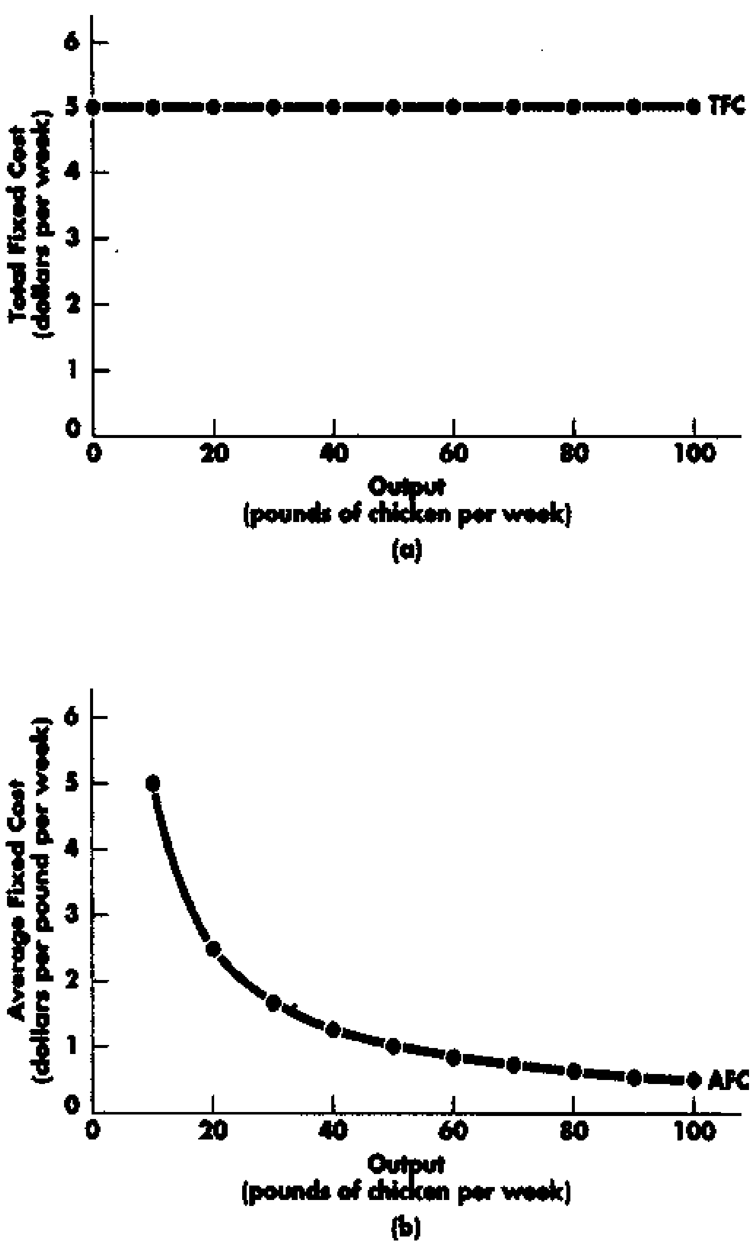
[3]Note, however, that an increase in traffic will increase annual maintenance and replacement cost, so replacement and maintenance are variable costs. Note also that opportunity costs can be fixed, variable, or a combination of the two.

[4]Here we assume that the fixed costs are also what economists call *sunk costs*, meaning that the firm has already spent the money in question or signed a contract to do so. Consequently, if the firm decides to go out of business (produce zero output), it must still spend the money.

FIGURE 7-5

FIXED COSTS: TOTAL AND AVERAGE

The total fixed cost curve in Panel (a) is horizontal because, by definition, TFC does not change when output changes. AFC decreases steadily in Panel (b) as the TFC is spread among more and more units of output; because AFC never reaches zero, however, the AFC curve never crosses the horizontal axis.



Average fixed cost, however, gets smaller and smaller as output increases because AFC (which equals TFC/Q) falls as output (the denominator) rises for constant TFC. Business people typically put the point another way: Any increase in output spreads the fixed cost among more units, leaving less of it to be carried by any one unit. For example, suppose that Flo's farm incurs a fixed cost of $5 per week. When she produces only one 20-pound package of chicken, the entire $5 of fixed cost must be borne by that 1 unit of output. But if she produces two 20-pound packages, each of them must cover only half of the total—$2.50 each (Table 7-4).

However, AFC can never reach zero. Even if the farm were to produce, say, 1 million packages of chicken per week, each unit would have to bear, on the average, one millionth of the TFC—which is still a positive number (though very

TABLE 7-4

TOTAL AND AVERAGE FIXED COSTS ON FLO'S CHICKEN FARM

Output per Period (20-lb. packages per week)	Total Fixed Cost (TFC)	Average Fixed Cost (AFC)
0	$5	—
1	5	$5.0
2	5	2.5
3	5	1.7
4	5	1.3
5	5	1.0
6	5	0.8
7	5	0.7
8	5	0.6

small). It follows that the AFC curve gets lower and lower as output increases, moving closer and closer to the horizontal axis, but never crossing it. This is the pattern shown in Figure 7-5(b).

The total fixed cost curve is always horizontal because, by definition, total fixed cost does not change when output changes. The average fixed cost curve declines when output increases, moving closer and closer to the horizontal axis but never crossing it.

Having divided costs into two parts, fixed costs (FC) and variable costs (VC), we can express corresponding and rather obvious rules for total and average costs:[5]

$$TC = TFC + TVC \qquad AC = AFC + AVC$$

SHAPE OF THE AVERAGE COST CURVE

The preceding discussion of fixed and variable costs enables us to complete our consideration of the shape of the average cost curve. Curve AC in Figure 7-4(b) on page 155 is *U*-shaped: the left-hand portion of the curve is downward sloping and the right-hand portion is upward sloping. AC declines when output increases in the left-hand portion of the curve for two reasons.

The first is the role of changing input proportions as the firm increases the quantity of one input while holding the others constant. If Flo is using very little corn relative to the amounts of other inputs, a rise in the quantity of corn can at first yield increasing additions to output (in the range of increasing marginal physical product of corn illustrated in the left-hand part of Figure 7-3). As the quantity increases, this tends to reduce the average cost of output.

The second reason why the left-hand portion of the firm's average cost curve tends to decline pertains to fixed costs. As Figure 7-5(b) shows, the average *fixed-*

[5]The reader may wonder if there is such a thing as a marginal fixed cost. The answer is *yes*, but it doesn't matter because marginal fixed cost must generally equal zero. By definition, an increase in output never adds anything to total fixed cost. For example, for both 2 and 3 units of output, TFC must be the same, so that $MFC_3 = TFC_3 - TFC_2 = 0$.

cost curve always falls as output increases, and it falls very sharply at the left-hand end of curve AFC. But since AC equals AFC plus AVC, the AC curve for virtually any product contains a fixed-cost portion, AFC, which falls steeply at first when output increases. That is the main reason that the AC curve for any product should have a downward-sloping portion such as *CD* in Figure 7-4(b), a portion which is said to be characterized by decreasing average cost.

To the right of point *E* in the same figure, AC rises as output increases. This means that a given percentage rise in output requires a greater percentage rise in TC, so that AC = TC/Q must rise. Why does the portion of the curve with decreasing AC come to an end? The curve begins to turn upward for two reasons: (1) the law of diminishing returns, and (2) the administrative (bureaucratic) problems of large organizations.

The first of these phenomena is central to our present discussion, for we are expanding one input (the quantity of corn) while holding all others constant. In that case, the law of diminishing returns will surely work to increase both marginal and average costs, for reasons already discussed.

But the second, and probably more important, source of increasing average cost in practice is a matter of sheer size. Large firms tend to be relatively bureaucratic, impersonal, and costly to manage. As a firm becomes very large and loses the personal touch of top management, bureaucratic costs will ultimately rise disproportionately. This change ultimately drives average cost upward.

The point at which average cost begins to rise varies from industry to industry. It occurs at a much larger volume of output in automobile production than in farming—which is why no farms are as big as even the smallest auto producer. A large part of the reason is that the fixed costs of automobile production are far greater than those in farming, so spreading the fixed cost over an increasing number of units of output keeps AC falling far longer in auto production than in farming. Thus, although firms may have *U*-shaped AC curves in both industries, the bottom of the *U* occurs at a far larger output in auto production than in farming.

A **SUNK COST** is a cost to which a firm is precommitted for some limited period, either because it has signed a contract to make the payments or because it has already paid for some durable item (such as a machine or a factory) and cannot get its money back except by using that item to produce output for some period of time.

The AC curve for a typical firm is *U*-shaped. Its downward-sloping segment is attributable to increasing marginal returns and/or to the fact that the firm spreads its fixed costs over larger and larger outputs. The upward-sloping segment is attributable to decreasing marginal returns and to the disproportionate rise in administrative cost that occurs as firms grow large. The output at which decreasing average cost ends and increasing average cost begins varies from industry to industry. Other things being equal, the greater the relative size of fixed costs, the higher will be the output at which the switchover occurs.[6]

LONG-RUN VERSUS SHORT-RUN COSTS

The cost to the firm of a change in its output depends very much on the period of time under consideration. At any point in time, many input choices are *precommitted* by past decisions. If, for example, a firm purchased machinery a year ago, it is committed to that decision for the remainder of the machine's economic life, unless the company is willing to take the loss involved in getting rid of it sooner. The cost is then said to be a **sunk cost.**

[6]Empirical evidence confirms this view, though it suggests that the bottom of the *U* is often long and flat. That is to say, a considerable range of outputs often fall between the regions of decreasing and increasing average cost. In this intermediate region, the AC curve is approximately horizontal, meaning that, in that range, AC does not change when output increases.

Whether or not a cost is sunk depends on the planning horizon of a particular decision. For example, a 2-year-old machine with a 9-year economic life is an inescapable commitment, and therefore a sunk cost, for the next 7 years. But it is not an unchangeable commitment (sunk cost) in plans that extend beyond 7 years, for the machine will have to be replaced in any case. Economists summarize this notion by speaking of two different "runs" for decision making—the **short run** and the **long run.**

The **SHORT RUN** is a cost period of time over which some of the firm's cost commitments will have ended.

The **LONG RUN** is a period of time long enough for all of the firm's sunk commitments to come to an end.

These terms will recur time and again in this book. They emerge now because of their relationship to the shape of the cost curve. In the short run, the firm has relatively little opportunity to adapt its production processes to the size of its current output because the size of its plant has largely been predetermined by its past decisions. Over the long run, however, all inputs, including the size of the plant, become adjustable.

Consider the example of Flo's Farm. Once she has constructed her chicken coop, she has relatively little discretion over her production capacity. She can build an addition, but that may be considerably more expensive than building a bigger coop in the first place. Cutting the size of the coop is even more costly. Over a somewhat longer planning horizon, however, the original chicken coop will need to be replaced, and Flo will be free to decide all over again how large a building to construct.

Much the same is true of big industrial firms. Companies have little control over their plant and equipment capacities in the short run. But with some advance planning, they can acquire different types of machines, redesign factories, and make other choices. For instance, General Motors continued producing the Chevrolet Caprice and other big, rear-drive cars at its plant in Arlington, Texas, for the 1995 and 1996 model years even though the vehicles were not selling well. That was partly because the company knew that it would need time to convert the plant to manufacture its popular full-size pickup trucks, which were in short supply. By the 1997 model year, however, GM engineers were able to convert the plant to truck production.

It should be noted that the short run and long run do not refer to the same period of time for all firms; rather, they vary in length, depending on the nature of the firm's sunk commitments. If, for example, the firm can change its work force every week, its machines every 2 years, and its factory every 20 years, then 20 years will be the long run, and any period less than 20 years will constitute the short run.

THE AVERAGE COST CURVE IN THE SHORT AND LONG RUN

As we just observed, which inputs can be varied and which are precommitted depends on the pertinent time horizon. It follows that:

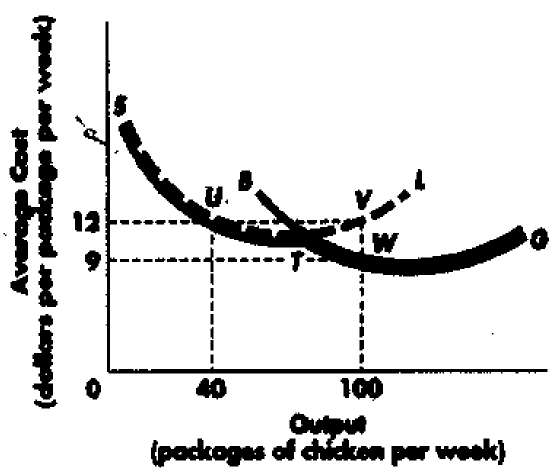
The average (and marginal and total) cost curve depends on the firm's planning horizon. The average (and total) cost curve for the long run differs from that for the short run because in the long run more inputs become variable.

We can, in fact, be much more specific about the relationships between the short-run and long-run average cost (AC) curves. Consider, as an example, the capacity of Flo's chicken farm. In the short run, after she has built her chicken coop, she can choose only the number of chickens to crowd into its capacity. But, in the long run, she can also choose among buildings of different sizes.

If she constructs a smaller building, her AC curve looks like curve *SL* in Figure 7-6. That means that if she is pleasantly surprised and sales grow to 100

FIGURE 7-6 SHORT-RUN AND LONG-RUN AVERAGE COST CURVES

Flo has a choice of two sizes of chicken coop, a small one with AC curve *SL*, and a big one with AC curve *BG*. These are the short-run curves that apply as long as the farm is stuck with its chosen coop. In the long run, however, when it has its choice of coop size, it can pick any point on the orange lower boundary of these curves. This lower boundary, *STG*, is the long-run average cost curve.



pounds per week, average cost will be $12 per 20-pound package per week (point *V*). She may then wish she had built the bigger chicken coop with an AC curve of *BG*, which would have enabled her to cut unit cost to $9 (point *W*). However, in the short run, Flo can do nothing about this decision; the AC curve remains *SL*. Similarly, had she built the larger chicken coop, the short-run AC curve would be *BG*, and the farm would be committed to this cost curve even if business were to decline sharply.

In the long run, however, Flo must replace the building, and she is free to decide all over again how big a coop to construct. If Flo expects sales of 100 pounds per week, she will construct the larger building and have an average cost of $9 per package of chicken. Similarly, if she expects sales of only 40 packages, she will arrange for the smaller building with average costs of $12 (point *U*).

In sum, in the long run, a firm will select the plant size (that is, the short-run AC curve) that is most economical for the output level that it expects to produce. The long-run average cost curve then consists of all of the *lower* segments of the short-run AC curves. In Figure 7-6, this composite curve is the orange curve, *STG*.

The long-run average cost curve shows the lowest possible short-run average cost corresponding to each output level.

MULTIPLE INPUT DECISIONS: THE CHOICE OF OPTIMAL INPUT COMBINATIONS

Up to this point in the chapter we have simplified the analysis by assuming that the firm can change the quantity of only one of its inputs. We come next to the more realistic case in which the firm can and does decide on the quantities of

[7] Instructors may want to teach this part of the chapter (up to page 167) now, or they may prefer to wait until they come to Chapters 15 and 16 on the determination of wages, interest rates, profit, and rent.

several inputs—such as labor, land, and capital—at once. We will once again examine the three basic and closely interrelated subjects—production, optimal choice of input quantities, and the costs of the firm—but this time we will allow the firm to select the quantities of *many* inputs. By expanding our analysis in this way, we can study a key issue: how a firm, by its choice of production method, can make up for decreased use of one input by using more of another. We will end by returning to the determination of the three cost curves (total cost, average cost, and marginal cost) that are so fundamental for understanding how the firm's pricing and output decisions depend upon its production relationships—that is, the connection between input and output quantities.

SUBSTITUTABILITY: THE CHOICE OF INPUT PROPORTIONS

Many people mistakenly believe that management really has very little choice when selecting its input proportions. Technological considerations alone, it would appear, dictate such choices. A particular type of furniture-cutting machine may require two operators working for an hour on a certain amount of wood to make five desks—no more and no less. But this is an overly narrow view of the matter. Whoever first said that there are many ways to skin a cat saw things more clearly.

The furniture manufacturer may choose among several alternative production processes for making desks. For example, simpler and cheaper machines may be able to change the same pile of wood into 5 desks using more than 2 hours of labor. Still more workers could eventually do the job with simple hand tools, using no machinery at all. The firm will seek the method of production that is *least costly.*

In advanced industrial societies, where labor is expensive and machinery is cheap, it may pay to use the most automated process. For example, Caterpillar, an American producer, curbed its high labor costs by investing in computers that enabled it to manufacture twice as many truck engines with the same number of people. But in less developed countries, where machinery is scarce and labor abundant, making things by hand may be the most economical solution. An interesting example can be found in India, where company records are often still handwritten, in contrast to American firms where records are computerized.

We conclude that one input can generally be *substituted* for another. A firm can produce the same number of desks with less labor, *if* it is prepared to sink more money into machinery. But whether it *pays* to make such a substitution depends on the relative costs of labor and machinery. Several general conclusions follow from this discussion.

1. Normally, a firm can choose among different options to produce a particular volume of output. Input proportions are rarely fixed immutably by technological considerations.
2. Given a target level of production, a firm that cuts down on the use of one input (say, labor) will normally have to increase its use of another input (say, machinery). This is what we mean when we speak of *substituting* one input for another.
3. The combination of inputs that represents the *least costly* way to produce the desired level of output depends on the relative prices of the various inputs.

In this section, we will discuss how a firm can determine the most economical input combination. This analysis is not just applicable to business enterprises.

Nonprofit organizations like your own college are interested in finding the least costly ways to accomplish a variety of tasks (for example, maintaining the grounds and buildings); government agencies are concerned with meeting their objectives at minimum costs; even in the home, there are many ways to "skin a cat." Thus, our present analysis of *cost minimization* is widely applicable.

THE MARGINAL RULE FOR OPTIMAL INPUT PROPORTIONS

Choosing the input proportions that minimize the cost of producing a given output is really a matter of common sense. To understand why, let us turn, once again, to our concrete example. Suppose that Florence Farmer is considering whether to feed her flock more corn and less soy meal, or vice versa. The two feeds are substitutes; if the chickens get more soy meal, they need less corn. *But the feeds are not perfect substitutes.* Soy meal provides more protein but fewer carbohydrates than corn, so Flo gains a considerable benefit by providing the chickens with a balance of the two. If their diet contains too much of one feed and too little of the other, the output of chicken meat will suffer. In other words, it is reasonable to assume *diminishing returns* to excessive substitution of either type of feed for the other.

Now suppose that corn costs $10 per 40-pound bag, as in our earlier example, while the same size bag of soy meal costs twice as much. Given the $10 and $20 prices for corn and soy, how much of each feed should Flo use? To answer this question, she must compare the prices of each feed with information on what each of them yields—their marginal physical products. If the marginal physical product (MPP) of a bag of corn is 30 pounds of meat, but the marginal physical product of an additional bag of soy meal is 50 pounds, what should Flo do? A little thought reveals the answer: Flo should cut down on soy meal and increase her use of corn. Why? Because soy meal costs twice as much as corn, but yields only 67 percent more meat. Let us examine the reasoning a bit more carefully.

To see why the farmer can produce the same output of chicken at a lower input cost by spending less on soy meal and more on corn, suppose that Flo buys one less bag of soy meal, thus saving $20. This reduces her output by 50 pounds of chicken—the marginal product of soy meal. How much more corn does she need to replace the lost output? Since the MPP of corn is 30 pounds, it will take 50/30, or 1.67 additional bags of corn to make up the shortfall. But, at a price of $10 per bag, the 1.67 bags of corn will cost Flo just $16.70. Thus, by using one less bag of soy meal and 1.67 more bags of corn, she ends up saving $20.00 − $16.70 = $3.30 and producing the same output—a good deal!

What makes this example work out the way it does is that *the ratio of the marginal product of soy meal to the price of soy meal is less than the ratio of the marginal product of corn to the price of corn.* Specifically, the MPP of soy meal is 50 pounds and its price is $20, so the MPP per dollar is 50 pounds/$20, or 2.5 pounds per dollar. By contrast, the MPP per dollar of expenditure on corn is 30 pounds/$10, or 3.0 pounds per dollar, which is larger. This means that farmers get more for their money—at the margin—by spending on corn rather than on soy meal.

To generalize, a firm can interpret the ratio:

MPP of any input/Price of that input

as the marginal product of spending $1 on the input. It should be fairly clear that it always pays to reduce spending on Input A if its MPP per dollar is less than the MPP per dollar of Input B. That dollar should be spent on *B* instead. Such a

move must *always* reduce the firm's costs. By switching away from the input with the *lower* marginal product per dollar, and buying more of the input with the *higher* marginal product per dollar, the firm can reduce the money it spends without reducing its output.

We have thus derived the basic rule for determining the most economical way to produce any given output:

A firm can reduce the cost of producing a given output by using less of some input, A, and making up for it by using more of another input, B, whenever the ratio of the marginal physical product of A to the price of A is less than the ratio of the marginal physical product of B to the price of B, that is, whenever:

$$MPP_a/P_a < MPP_b/P_b$$

The opposite will be true if the MPP per dollar spent on A is higher than the MPP per dollar spent on B. Putting these two together, we conclude that input proportions *cannot* be optimal if the ratios of the marginal physical products to the prices for any two inputs differ, that is:

The proportions of any two inputs, A and B, used by the firm can be optimal only if:

$$MPP_a/P_a = MPP_b/P_b$$

This rule, as we have noted, is simply common sense. If a unit of Input A has a marginal product that is, say, three times as big as that of Input B, the firm should be willing to pay exactly three times as much money for an additional unit of A as it pays for a unit of B, no more and no less.

But what if the market happens to set input prices so that this doesn't work? Suppose that the market price of A happens to be twice as large as that of B, as in our chicken-growing example. What can the farmer do about it? In this case, the farmer should buy less of A (soy meal, in our example) and more of B (corn). That will not change the market *prices* of corn and soy meal, but it will change the *marginal physical products* of the two inputs because of the law of diminishing returns. As Flo buys more corn and less soy meal, the marginal product of corn will fall and the marginal product of soy meal will rise. When Flo has gone just far enough in switching money from soy meal to corn, the ratio of their marginal products will rise from 50:30 (1.67) up to, say, 56:28 (2.00). At that point, the ratio of the prices of the two inputs will be equal to the ratio of their marginal products, satisfying the rule for cost minimization.

CHANGES IN INPUT PRICES AND OPTIMAL INPUT PROPORTIONS

The common-sense reasoning behind the rule for optimal input proportions leads to an important conclusion. Suppose Flo is producing 120 pounds of output at minimum cost. Then the price of corn rises while the price of soy meal remains the same. That will cause a violation of the optimal-purchase rule:

$$MPP_{corn}/P_{corn} = MPP_{soy\ meal}/P_{soy\ meal}$$

To restore optimality the MPP of corn must also rise to match the rise in corn price. But, by the "law" of diminishing returns, the MPP of corn is *higher* only when the use of corn is lower. Thus, a rise in the price of corn leads the farmer

INPUT SUBSTITUTION IN THE FOREST

After logging was outlawed in much of the Redwood National Forest, Louisiana Pacific Corporation (one of the world's largest lumber producers) lost access to one of its essential inputs: old-growth timber from which the company made building materials. But CEO Harry Merlo was pragmatic about the situation, and turned his attention to trees that were not under the watchful eye of conservationists. One likely candidate, the aspen—a so-called *waste-species*—immediately came to his attention. Aspens grow quickly and forests stretching the width of North America contain plenty of them. Fortunately, Merlo had kept himself abreast of the latest technology in the building industry.

Canadian lumber mills had long had the capability of trimming small trees into sheets and then gluing and pressing the sheets together. Merlo had observed this process, and he had conceived some innovative ideas of his own. He correctly figured that alternating the grain of the layered sheets would add to overall strength. The result was a product, waferwood, that turned out to be cheaper and stronger than plywood. It could be used as siding and, when combined with other engineered lumber, used to create I-beams for floor joists and roof rafters.

Thus, when Louisiana Pacific was, in effect, priced out of the market for its old input (old-growth timber), it was led to substitute a new input, aspen wood, for one that was no longer available.

SOURCE: "Let's Go for Growth," *Fortune*, March 7, 1994, p. 60.



to use *less* corn and, if she still wants to produce 120 pounds of chicken, to use *more* soy meal. In general:

As any one input becomes more costly relative to competing inputs, the firm is likely to substitute one input for another; that is, to reduce its use of the input that has become more expensive and to increase its use of competing inputs.

This general principle of input substitution applies in industry just as it does on Flo's farm. For an unusual application of the analysis, see the box above.

THE PRODUCTION FUNCTION AND THE FIRM'S COST CURVES

The **PRODUCTION FUNCTION** indicates the *maximum* amount of product that any particular collection of inputs is capable of producing.

To help select the combination of inputs that can produce the desired output at least cost, economists have invented a concept they call the **production function.** The production function summarizes the technical and engineering information about the relationship between a firm's output and *all* of its inputs. It indicates, for example, just how much chicken meat Flo can produce with given amounts of land, labor, corn, and so on. If, as we normally assume, the firm can produce the desired amount of output in many ways, the production function lays out all of the input combinations that will do the job.

TABLE 7-5 ALTERNATIVE INPUT COMBINATIONS FOR PRODUCTION OF GIVEN OUTPUT QUANTITIES

Ways to Produce 80 Pounds of Chicken per Week			Ways to Produce 120 Pounds of Chicken per Week		
Corn (bags per week)	Soy Meal (bags per week)	Total Cost[a] (per week)	Corn (bags per week)	Soy Meal (bags per week)	Total Cost[a] (per week)
5.0	0	$50	7.0	0	$ 70
3.5	0.5	45	4.5	0.7	59
2.0	1.0	40	2.9	1.4	57
1.5	2.1	57	2.0	2.8	76
0	4.1	80	0	5.5	110

[a]Total cost is calculated from the formula TC = Price of corn × Quantity of corn + Price of soy meal × Quantity of soy meal, with corn priced at $10 per bag and soy meal at $20 per bag.

If a firm needs only two inputs—which are enough to indicate the basic principles—a production function can be represented graphically (which we do in the appendix to this chapter) or by a simple table. Table 7-5 shows part of Flo's production function for the use of soy meal and corn to produce chicken meat. It lists some of the combinations of soy meal and corn that are capable of yielding 80 pounds of chicken per week, as well as the input combinations that yield 120 pounds. For example, if Flo wants to produce 120 pounds, she can examine the menu of choices offered by the right-hand panel.

How much labor and corn should Flo use to produce 120 pounds of chicken? The production function table shows a variety of options. She can, for example, use 7 bags of corn per week and no soy meal or 4.5 bags of corn and 0.7 bags of soy meal. At the other extreme, she can raise the 120 pounds of chicken without any corn by using 5.5 bags of soy meal.

Which will she choose? Naturally, she prefers the one that costs the least. The least-cost option can be found directly with the help of the table and the information that corn costs $10 per bag while soy meal is priced at $20. The blue screen indicates the lowest entry in the last column—$57—corresponding to the use of 2.9 bags of corn and 1.4 bags of soy meal. That implies that this input combination is the most economical way to produce the 120 pounds of chicken. The last column in the table also shows the total cost of each of the other input combinations capable of producing 120 pounds of chicken. (Note that each total cost entry is calculated, as before, by multiplying the input quantities by the input prices and adding them together; for the least-cost option, 2.9 × $10 + 1.4 × $20 = $57.) Similarly, the third column in the table indicates that 80 pounds of chicken can be produced most inexpensively using 2 bags of corn and 1 bag of soy meal, costing $40 ($10 × 2 + $20 × 1 = $40).

Outputs of 80 and 120 pounds per week are not the only production levels that Flo should consider. She also should consider, for example, the least costly way to produce 60 pounds, 180 pounds, and so on. The total cost of producing *any* given quantity of output equals the sum of the cost-minimizing quantities of each of the inputs, with each input quantity multiplied by the price of that input.

Using the procedures just outlined, Flo can calculate the minimum cost of producing *any* quantity of output. Let us now suppose that she has done this,

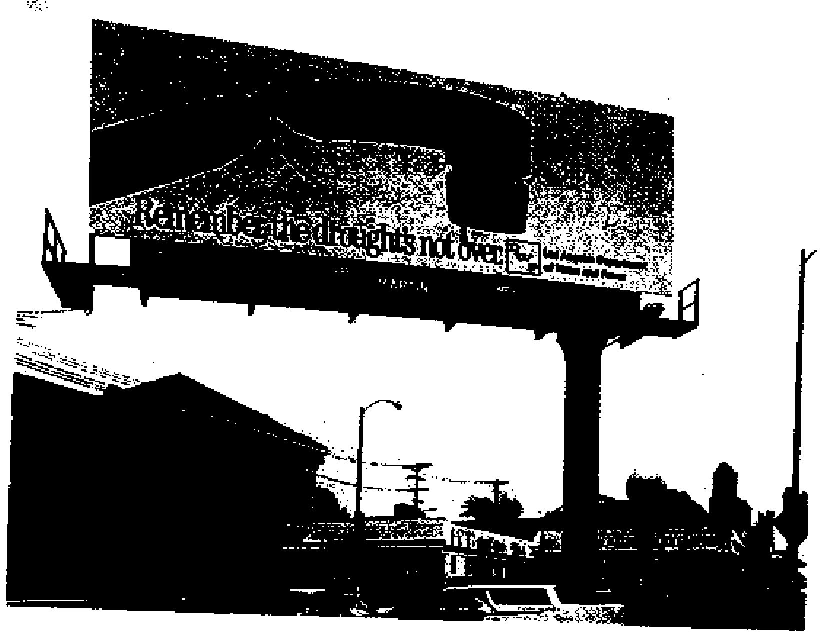
POLICY DEBATE **SHOULD WATER BE PROVIDED TO WESTERN FARMERS AT SUBSIDIZED PRICES?**

Western farmers use a great deal of water. It is one of the critical inputs to their production processes. Yet, for all practical purposes, it is provided to them virtually free. For example, government controls keep the price of water used for agriculture artificially low, so that California farmers pay only about one-thirtieth of the price that urban residents pay for water. Even during a recent drought, farmers in that state continued to use vast quantities of water, while residents in the cities were forced to ration.

This situation has given rise to a spirited debate between farmers and environmentalists. There is no question that water is scarce in the West, and some predictions foreshadow a looming shortage of disastrous proportions. It is also clear that farmers pay a price that is much lower than the true marginal cost of water, particularly because that cost includes a very high *opportunity cost*, that is, the value of the other uses of water that must be forgone as a result of its extensive employment in agriculture.

As the analysis of this chapter shows, a low price for an input increases the amount that is used, and there is little doubt that the low price of water substantially increases its use by western farmers. Environmentalists and economists have joined forces in arguing that users of water should pay prices that cover its true marginal cost. Indeed, it has been suggested that at such a price any shortage would simply disappear.

The farmers respond that long practice entitles them to continued low water prices and that low prices in the past induced them to invest extensively in their agricultural properties, so that a price increase now would be tantamount to confiscation of their investments. Moreover, they argue that a sharp rise in the price of this critical input would destroy a way of life that is valuable for society. The politicians undoubtedly feel caught in the middle, as the debate continues.



completing the information that she needs to plot her *total cost curve*. As before, dividing the total cost for each output by the quantity of output then gives the corresponding *average cost*, that is, the cost per unit of output. Similarly, she can derive the marginal cost curve from the total cost figures, just as an earlier section did.

ECONOMIES OF SCALE

Production is said to involve **ECONOMIES OF SCALE**, also referred to as **INCREASING RETURNS TO SCALE**, if, when all input quantities are doubled, the quantity of output is more than doubled.

We are now beginning to put together the tools we need to address the question posed at the beginning of this chapter: Does a large firm benefit from substantial **economies of scale** that allow it to operate more efficiently than smaller firms? To answer this question, we need a precise definition of this concept.

The scale of operation of an enterprise is defined by the quantities of the various inputs that it uses. The production function illustrates what happens when the firm doubles its scale of operations. For an example of economies of scale,

turn back to Flo's production function in Table 7-5, and assume that soy meal and corn are the only two inputs.[8] Notice that with 1.5 bags of corn and 2.1 bags of soy meal, output is 80 pounds of chicken. But if Flo increases both inputs by 33.33 percent—to 2 bags of corn and 2.8 bags of soy meal—the table shows that output rises by 50 percent (to 120 pounds). Since output goes up by a greater percentage than all of the inputs, Flo's production function displays **increasing returns to scale** (also known as economies of scale), at least in this range.

Economies of scale affect operations in many modern industries. Where they exist, they give larger firms cost advantages over smaller ones and thereby foster large firm sizes. Automobile production and telecommunications are two common examples of industries with important economies of scale. Predictably, firms in these industries are, indeed, huge.

Technology generally determines whether or not a specific economic activity is characterized by economies of scale. One particularly clear example is warehouse space. Imagine two warehouses, each shaped like perfect cubes with the length, width, and height of Warehouse 2 twice as large as those of Warehouse 1. Now remember your high-school geometry. The surface area of any side of a cube is equal to the square of its length. Therefore, the amount of material needed to build Warehouse 2 will be 2^2, or 4 times as great as that needed for Warehouse 1. However, since the volume of a cube is equal to the cube of its length, Warehouse 2 will have 2^3, or 8 times as much storage space as Warehouse 1. Thus, multiplying the inputs by 4 leads to 8 times the storage space—an example of strongly increasing returns to scale.

This example is, of course, oversimplified. It omits such complications as the need for stronger supports in taller buildings, the increased difficulty of moving goods in and out of higher stories, and the like. Still, the basic idea is correct, and the example shows why, up to a point, the very nature of warehousing creates technological relationships that lead to economies of scale.

Our definition of economies of scale, though based on the production function, is closely related to the shape of the *long-run* average cost curve. Notice that the definition requires that a doubling of *every* input must bring about more than a doubling of output. If all input quantities are doubled, total cost must double. But if output *more* than doubles, then cost per unit (average cost) must decline. In other words:

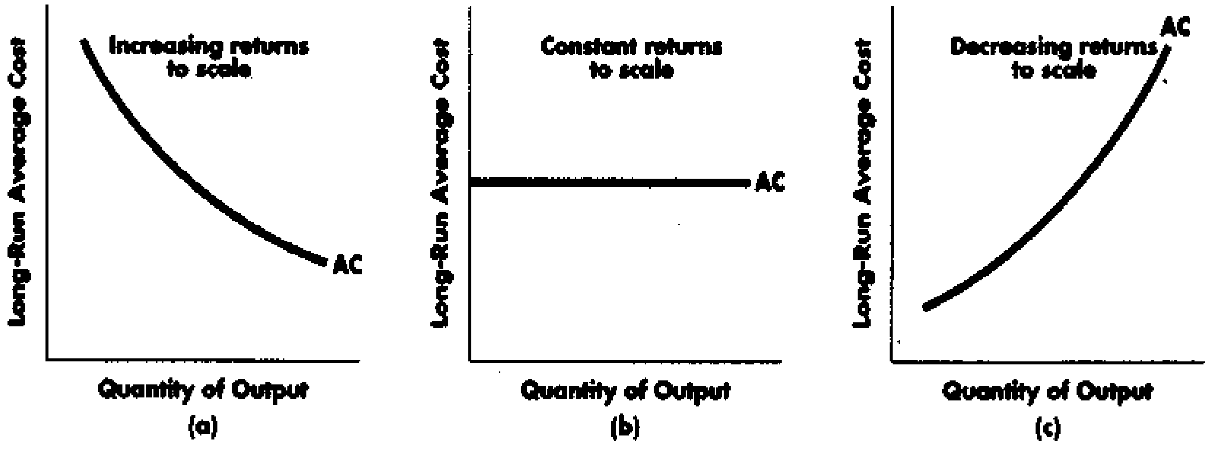
Production functions with economies of scale lead to long-run average cost curves that decline as output expands.

Figure 7-7(a) depicts a decreasing average cost curve, but this is only one of three possible shapes that the long-run average cost curve can take. Panel (b) shows the curve for constant returns to scale. Here, if all input quantities double, both total cost (TC) and the quantity of output (Q) double, so average cost ($AC = TC/Q$) remains *constant*. Finally, output may increase, but less than double, when all inputs double. This is a case of decreasing returns to scale, which leads to a *rising* long-run average cost curve like the one depicted in Panel (c). The figure reveals a close association between the slope of the AC curve and the nature of the firm's returns to scale.

It should be pointed out that the same production function can display increasing returns to scale in some ranges, constant returns to scale in other

[8]This is necessary because the table deals with only two inputs and the definition requires that *all* inputs be doubled simultaneously. To be true to the definition, because labor, corn, *land, and machinery* were all used by the farmer, their quantities would all have to double.

FIGURE 7-7 THREE POSSIBLE SHAPES FOR THE LONG-RUN AVERAGE COST CURVE



In Panel (a), long-run average costs are decreasing as output expands because the firm has significant economies of scale (increasing returns to scale). In Panel (b), constant returns to scale lead to a long-run AC curve that is flat; costs per unit are the same for any level of output. In Panel (c), which pertains to a firm with decreasing returns to scale, long-run average costs rise as output expands.

ranges, and decreasing returns to scale in yet others. This is true of all the *U*-shaped average cost curves we have shown, for example, Figure 7-4(b).

THE "LAW" OF DIMINISHING RETURNS AND RETURNS TO SCALE

Earlier in this chapter, we discussed the "law" of diminishing marginal returns. Is there any relationship between economies of scale and the phenomenon of diminishing returns? It may seem at first that the two are contradictory. After all, if a producer gets diminishing returns from his inputs as he uses more of each of them, doesn't it follow that by using more of *every* input, he must encounter decreasing returns to scale? In fact, the two principles do not contradict one another, for they deal with fundamentally different issues.

1. *Returns to a single input.* This analysis asks the question: How much does output expand if a firm increases the quantity of just *one* input, *holding all other input quantities unchanged*?
2. *Returns to scale.* Here the question is: How much does output expand if *all* inputs are increased *simultaneously* by the same percentage?

The "law" of diminishing returns pertains to the first question since it examines the effects of increasing only one input at a time. It is plausible that the firm will encounter diminishing returns as this one input becomes relatively abundant. Thus, for example, the addition of too much soy meal relative to a given quantity of corn will contribute too much protein and too few carbohydrates to the chickens, yielding diminishing returns.

Returns to scale pertain to proportionate increases in *all* inputs, and therefore answer the second question. If Flo doubles the quantities of both corn and soy meal, the chickens' diet need not become unbalanced. Thus, the "law" of diminishing returns (to a single input) is compatible with *any* sort of returns to scale. In summary:

Returns to scale and returns to a single input (holding all other inputs constant) refer to two distinct aspects of a firm's technology. A production function that displays diminishing returns to *a single input* may show diminishing, constant, or increasing returns when *all input quantities are increased proportionately*.

HISTORICAL COSTS VERSUS ANALYTICAL COST CURVES

In Chapter 6, we noted that all points on a demand curve pertain to the *same* period of time. Decision makers must use this common time period because the demand curve illustrates alternative choices *for a given period of time*. The same is true of a cost curve. All points on a cost curve pertain to exactly the same time period because the graph examines the cost of each alternative output level that the firm could choose for that period.

It follows that a graph of historical data on prices and quantities at *different points in time* is normally *not* the cost curve that the decision maker needs. This observation helps to resolve the problem posed at the beginning of the chapter: Are declining historical costs evidence of economies of scale?

All points on any of the cost curves used in economic analysis refer to the same period of time.

One point on the cost curve of an auto manufacturer might show, for example, how much it would cost the firm to produce 2.5 million cars during 1997. Another point on the same curve might show what would happen to the firm's costs if, *instead*, it were to produce 3 million cars in 1997. Such a curve is called an *analytical cost curve* or, when there is no possibility of confusion, simply a cost curve. This curve must be distinguished from a diagram of *historical costs*, which shows how costs have changed from year to year.

The different points on an analytical cost curve represent *alternative possibilities*, all for the same time period. In 1997, the car manufacturer will produce either 2.5 million or 3 million cars (or some other amount), but certainly not both. Thus, at most, only one point on this cost curve will ever be observed. The company may, indeed, produce 2.5 million in 1997 and 3 million in 1998, but the 1998 data are not relevant to the 1997 cost curve. By the time 1998 comes around, the cost curve may have shifted, so the 1997 cost figure will not apply to the 1998 cost curve, either.

A different sort of graph can, of course, indicate year by year how costs and outputs have varied. Such a graph, which gathers together the statistics for a number of different periods, is not, however, a *cost curve* as that term is used by economists. An example of such a diagram of historical costs was presented in Figure 7-1 on page 148.

Why do economists rarely use historical cost diagrams and instead deal primarily with analytical cost curves, which are more abstract, harder to explain, and more difficult to estimate statistically? The answer is that analysis of real policy problems—such as the desirability of having a single supplier of telephone services for the entire market—leaves no choice in the matter. Rational decisions require analytical cost curves. Let us see why.

RESOLVING THE ECONOMIES OF SCALE PUZZLE

Recall the problem introduced early in the chapter. We examined the AT&T divestiture and concluded that in order to determine whether it made sense to

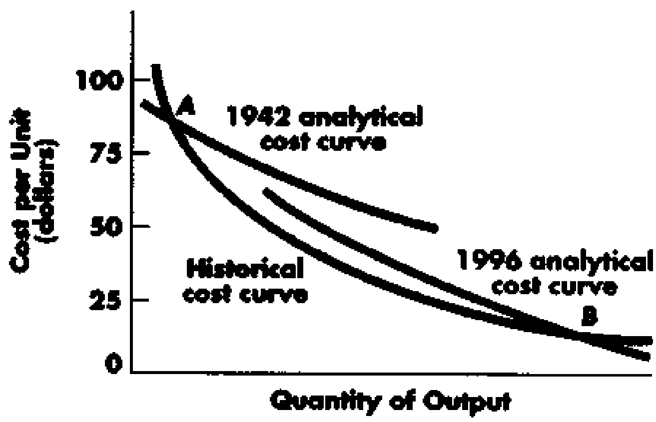
break up such a large company, economists would have to know whether the industry featured economies of scale. Among the data offered as evidence was a graph that showed a precipitous drop in the capital cost of long-distance communications as the volume of messages rose after 1942. The question left unanswered was why such information did not constitute legitimate evidence about the presence of economies of scale.

As this section shows, to determine whether a single supplier can provide telephone service more cheaply in 1996 than a number of smaller firms can, we must compare the costs of *both* large-scale and small-scale production *in 1996*. It does no good to compare the cost of a large supplier in 1996 with its own costs as a smaller firm back in 1942, because that cannot possibly supply the needed information. The cost situation in 1942 is irrelevant for today's decision between large and small suppliers because no small firm today would use the obsolete techniques of 1942.

Since the 1940s, great technical progress has taken the telephone industry from ordinary open-wire circuits to microwave systems, telecommunications satellites, coaxial cables of enormous capacity, and now to fiber optics. All of this means that the *entire* analytical cost curve of telecommunications must have shifted downward quite dramatically from year to year. Innovation must have reduced not only the cost of large-scale operations *but also the cost of smaller-scale operations*. Until decision makers compare the costs of large and small suppliers *today*, they cannot make a rational choice between single-firm and multifirm production. It is the analytical cost curve, all of whose points refer to the same period, that, by definition, supplies this information.

Figures 7-8 and 7-9 show two extreme hypothetical cases, one that entails true economies of scale and one that does not. Both are based on the same historical cost data (in black) with their very sharply declining costs. (This curve is reproduced from Figure 7-1.) They also show (in orange and blue) two possible average cost curves, one for 1942 and one for 1996. In Figure 7-8, the analytical AC curve has shifted downward very sharply from 1942 to 1996, as technological change reduced all costs. Moreover, both of the AC curves slope downward to the right, meaning that, in either year, a larger firm has lower average costs. Thus, the situation shown in Figure 7-8 really does entail scale economies, so that one large firm can serve the market at lower cost than many small ones.

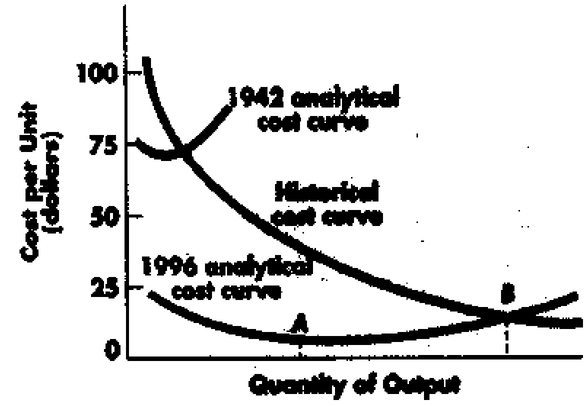
FIGURE 7-8 DECLINING HISTORICAL COST CURVE WITH THE ANALYTICAL AVERAGE COST CURVE ALSO DECLINING IN EACH YEAR



The two analytical cost curves shown indicate how the corresponding points (*A* and *B*) on the historical cost diagram are generated by that year's analytical curve. The declining analytical cost curves reveal economies of scale in the production activity.

FIGURE 7-9 DECLINING HISTORICAL COST CURVE WITH U-SHAPED ANALYTICAL COST CURVES IN EACH YEAR

Here the shape of the analytical average cost curves do not show economies of scale.



But now look at Figure 7-9, which shows exactly the same historical costs as Figure 7-8. Here, however, both analytical AC curves are *U*-shaped. In particular, the 1996 AC curve has its minimum point at an output level, *A*, that is less than one-half of the current output, *B*, of the large supplier. This means that, for the situation shown in Figure 7-9, a smaller company can produce more cheaply than a large one. In this case, one cannot justify domination of the market by a single large firm on the grounds that its costs are lower—despite the sharp downward trend of historical costs. In sum, the behavior of historical costs reveals nothing about the cost advantages or disadvantages of a single large firm. More generally:

Because a diagram of historical costs does not compare the costs of large and small firms *at the same point in time*, it cannot be used to determine whether an industry provides economies of large-scale production. Only the analytical cost curve can supply this information.

In the case of telephone service, some estimates indicate that economies of large-scale production do indeed exist. Presumably because of this, more than 10 years after the Bell telephone breakup, the typical firm providing long-distance telephone service is still very large, with AT&T, MCI, and Sprint in the vanguard. Yet, half a dozen or so other smaller firms still compete in this arena.

COST MINIMIZATION IN THEORY AND PRACTICE

Lest you be tempted to run out and open a business, confident that you now understand how to minimize costs, we should point out that decision making in business is a good deal harder than we have indicated here. Rare is the business executive who knows for sure what her production function looks like, or the exact shapes of her marginal physical product schedules, or the precise nature of her cost curves. No one can provide an instruction book for instant success in business. What we have presented here is, instead, a set of principles that constitutes a guide to good decision making.

Business management has been described as the art of making critical decisions on the basis of inadequate information, and our complex and ever-chang-

ing world often leaves people no alternative to educated guesses. Actual business decisions will at best approximate the cost-minimizing ideal outlined in this chapter. Certainly, practicing managers will make mistakes, but when they do their jobs well and the market system functions smoothly, the approximation may prove amazingly good. While no system is perfect, inducing firms to produce at the lowest possible cost is undoubtedly one of the jobs the market system does best.

SUMMARY

1. A firm's total cost curve shows its lowest possible cost of producing any given level of output. This curve is derived from the input combination that the firm uses to produce any given output and the prices of the inputs.
2. The **marginal physical product** of an input is the increase in total output resulting from a 1-unit increase in that input, holding the quantities of all other inputs constant.
3. The *"law" of diminishing marginal returns* states that if a firm increases the amount of one input (holding all other input quantities constant), the marginal physical product of the expanding input will eventually begin to decline.
4. To maximize profits, a firm must purchase an input up to the point at which diminishing returns reduce the input's **marginal revenue product** to equal its price.
5. The **long run** is a period sufficiently long for the firm's plant to require replacement and for all of its current contractual commitments to expire. The **short run** is any period briefer than that.
6. **Fixed costs** are costs whose total amounts do not vary when output increases. All other costs are called **variable costs.**
7. At all levels of output, the total fixed cost (TFC) curve is horizontal and the average fixed cost (AFC) curve declines toward the horizontal axis, but never crosses it.
8. TC = TFC + TVC; AC = AFC + AVC
9. It is usually possible to produce the same quantity of output in a variety of ways by substituting more of one input for less of another. Firms normally seek the least costly way to produce any given output.
10. A firm that wants to minimize costs will select input quantities at which the ratios of the **marginal physical product** of each input to its price—its MPP per dollar—are equal for all inputs.
11. The **production function** shows the relationship between inputs and output. It indicates the maximum quantity of output obtainable from any given combination of inputs.
12. If a doubling of all the firm's inputs *just* doubles its output, the firm is said to have *constant returns to scale.* If a doubling of all inputs leads to *more than* twice as much output, it has **increasing returns to scale** (or, **economies of scale**). If a doubling of inputs produces *less than* a doubling of output, the firm has *decreasing returns to scale.*
13. With increasing returns to scale, the firm's long-run average costs are decreasing; constant returns to scale are associated with constant long-run average costs; decreasing returns to scale are associated with increasing long-run average costs.
14. Economists cannot tell if an industry offers economies of scale (increasing returns to scale) simply by inspecting a diagram of historical cost data. Only the underlying analytical cost curve can supply this information.

KEY TERMS

Total physical product (TPP)
Average physical product (APP)
Marginal physical product (MPP)
Marginal revenue product (MRP)
Fixed cost
Variable cost
Sunk cost
Short run
Long run
Production function
Economies of scale (increasing returns to scale)

QUESTIONS FOR REVIEW

1. A firm's total fixed cost is $360,000. Construct a table of its total and average fixed costs for output levels varying from zero to 6 units. Draw the corresponding TFC and AFC curves.

2. With the following data, calculate the firm's AVC and MVC and draw the graphs for TVC, AVC, and MVC.

Quantity	Total Variable Costs (thousands of dollars)
1	$ 40
2	80
3	120
4	176
5	240
6	360

3. From the figures in Review Questions 1 and 2 above, calculate TC and AC for each of the output levels from 1 to 6 units and draw the two graphs.

4. If a firm's commitments in 1997 include machinery that will need replacement in 5 years, a factory building rented for 12 years, and a 3-year union contract specifying how many workers it must employ, when, from its point of view in 1997, does the firm's long run begin?

5. If the marginal revenue product of a gallon of oil used as input by a firm is $1.20 and the price of oil is $1.07 per gallon, what can the firm do to increase its profits?

6. A firm hires two workers and rents 15 acres of land for a season. It produces 150,000 bushels of crop. If it had doubled its land and labor, production would have been 325,000 bushels. Does it have constant, decreasing, or increasing returns to scale?

7. Suppose that wages are $20,000 per season per person and land rent per acre is $3,000. Calculate the average cost of 150,000 bushels and the average cost of 325,000 bushels, using the figures in Review Question 6. (Note that average costs increase when output increases.) What connection do these figures have with the firm's returns to scale?

8. Flo Farmer has stockpiled a great deal of corn. Suppose now that she buys more young chicks, but not more corn, and divides the corn she has evenly among the larger number of chicks. What is likely to happen to the marginal physical product of corn? What, therefore, is the role of input proportions in the determination of marginal physical product?

9. Labor costs $10 per hour. Nine workers produce 180 bushels of product per hour, while 10 workers produce 196 bushels. Land rents for $1,000 per acre per year. With 10 acres worked by nine workers, the marginal physical product of an acre of land is 1,400 bushels per year. Does the farmer minimize costs by hiring nine workers and renting 10 acres of land? If not, which input should he use in larger relative quantity?

10. Suppose that Flo's total costs increase by $12 per week at every output level. Show in Table 7-2 how this affects her total and average costs.

11. (More difficult) A firm experiences a sudden increase in the demand for its product. In the short run, it must operate longer hours and pay higher overtime wage rates to satisfy this new demand. In the long run, however, the firm can install more machines instead of operating fewer machines for longer hours. Which do you think will be lower, the short-run or the long-run average cost of the increased output? How is your answer affected by the fact that the long-run average cost includes the new machines the firm buys, while the short-run average cost includes no machine purchases?

APPENDIX

PRODUCTION INDIFFERENCE CURVES

To describe a production function—that is, the relationship between input combinations and the size of a firm's total output—economists can use a graphic device called the *production indifference curve* instead of the sort of numerical information described in Table 7-5 in the chapter.

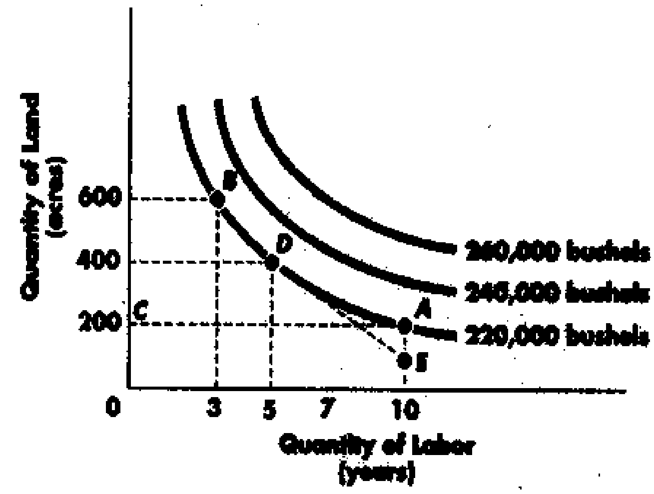
A **production indifference curve** (sometimes called an *isoquant*) is a curve in a graph showing quantities of *inputs* on its axes.

Each indifference curve indicates *all* combinations of input quantities capable of producing *a given* quantity of output; thus, a separate indifference curve corresponds to each possible quantity of output. These production indifference curves are perfectly analogous to the consumer indifference curves discussed in the appendix to Chapter 5.

Figure 7-10 represents different quantities of labor and capital capable of producing given amounts of wheat. The indifference curve labeled

FIGURE 7-10 A PRODUCTION INDIFFERENCE MAP

The figure shows three indifference curves, one for the production of 220,000 bushels of wheat, one for 240,000 bushels, and one for 260,000 bushels. For example, the lowest curve shows all combinations of land and labor capable of producing 220,000 bushels of wheat. Point *A* on that curve shows that 10 person-years of labor and 200 acres of land are enough to do the job.



220,000 bushels indicates that a farm can generate an output of 220,000 bushels of wheat using *any one* of the combinations of inputs represented by points on that curve. For example, it can employ 10 years of labor and 200 acres of land (point *A*) or the labor-capital combination shown by point *B* on the same curve. Because it lies considerably below and to the right of point *B*, point *A* represents a productive process that uses more labor and less land.

Points *A* and *B* can be considered *technologically* indifferent because each represents a bundle of inputs capable of yielding the same quantity of finished goods. However, the word "indifference" in this sense does not mean that the producer will be unable to decide between input combinations *A* and *B*. Input prices will permit the producer to arrive at a decision.

The production indifference curves in a diagram such as Figure 7-10 constitute a complete description of the production function. For each combination of inputs, they show how much output can be produced. Since production indifference curves are drawn in two dimensions, they represent only two inputs at a time. In more realistic situations, firms are likely to need more than two inputs, so managers must conduct an algebraic analysis. Still, all the principles we need to analyze such a situation can be derived from the two-variable case.

CHARACTERISTICS OF THE PRODUCTION INDIFFERENCE CURVES

Before discussing input pricing and quantity decisions, we first examine what is known about the shapes of production indifference curves.

Characteristic 1: *Higher curves correspond to larger outputs.* Points on a higher indifference curve represent larger quantities of *both* inputs than the corresponding points on a lower curve. Thus, a higher curve represents a larger *output*.

Characteristic 2: *An indifference curve will generally have a negative slope.* It goes downhill as we move toward the right. This means that if a firm reduces the quantity of one input, and it does not want to cut production, it must use more of another input.

Characteristic 3: *An indifference curve is typically assumed to curve inward toward the origin near its middle.* This shape reflects the "law" of diminishing returns to a single input. For example, in Figure 7-10, points *B*, *D*, and *A* represent three different input combinations capable of producing the same quantity of output. At point *B*, the firm uses a large amount of land and relatively little labor, while the opposite is true at point *A*. Point *D* is intermediate between the two.

Now consider the choice among these input combinations. When the farmer considers moving from point *B* to point *D*, he gives up 200 acres of land and instead hires 2 additional years of labor. Similarly, the move from *D* to *A* involves giving up another 200 acres of land. This time, however, hiring an additional 2 years of labor does not make up for the reduced use of land. Diminishing returns to labor as the farmer hires more and more workers to replace more and more land means that the farm now needs a much larger quantity of additional labor, 5 person-years rather than 2, to make up for the reduction in the use of land.

Without such diminishing returns, the indifference curve would have been a straight line, *DE*. The curvature of the indifference curve through points *D* and *A* reflects diminishing returns to substitution of inputs.

THE CHOICE OF INPUT COMBINATIONS

A production indifference curve describes only the input combinations that *can* produce a given output; it indicates the technological possibilities. In order to decide which of the available options suits its purposes best, a business needs the corresponding cost information: the relative prices of the inputs.

The budget line in Figure 7-11 represents equally costly input combinations for a firm. For example, if farmhands are paid $9,000 a year and land rents for $1,000 per acre a year, then a farmer who spends $360,000 can hire 40 farmhands but rent no land (point *K*), or he can rent 360 acres but have no money left for farmhands (point *J*). It is undoubtedly more sensible to pick some intermediate point on his budget line at which he divides the $360,000 between the two inputs.

If the prices of the inputs do not change, then the slope of the budget line for the $360,000 expenditure will be the same as that for $400,000 or for any other level of spending. For if the price of hiring a worker is 9 times as high as the cost of renting an acre, then the farmer can rent 9 times as many acres as he can hire farmhands with any given amount of money. Thus, with input prices given, the budget line for different amounts of expenditures will all be parallel, as in Figure 7-12.

A firm that is seeking to minimize costs does not necessarily have a fixed budget. Instead, it wants to produce a given quantity of output (say, 240,000 bushels) with the *smallest possible budget*.

Figure 7-12 combines the indifference curve for 240,000 bushels from Figure 7-10 with a variety of budget lines similar to *JK* in Figure 7-11. The firm's problem is to find the lowest budget line that will allow it to reach the 240,000-bushel indifference curve. Clearly, an expenditure of $270,000 is too little; no point on the budget line, *AB*, permits production of 240,000 bushels. Similarly, an expenditure of $450,000 is too much, because the firm can produce its target level of output more cheaply. The solution is at point *T* where the farmer uses 15 workers and 225 acres of land to produce the 240,000 bushels of wheat. In general:

The least costly way to produce any given level of output is indicated by the point of tangency between a budget line and the production indifference curve corresponding to that level of output.

COST MINIMIZATION, EXPANSION PATH, AND COST CURVES

Figure 7-12 shows how to determine the input combination that minimizes the cost of producing

FIGURE 7-11 A BUDGET LINE

The firm's budget line, *JK*, shows all of the combinations of inputs it can purchase with a fixed amount of money—in this case $360,000.

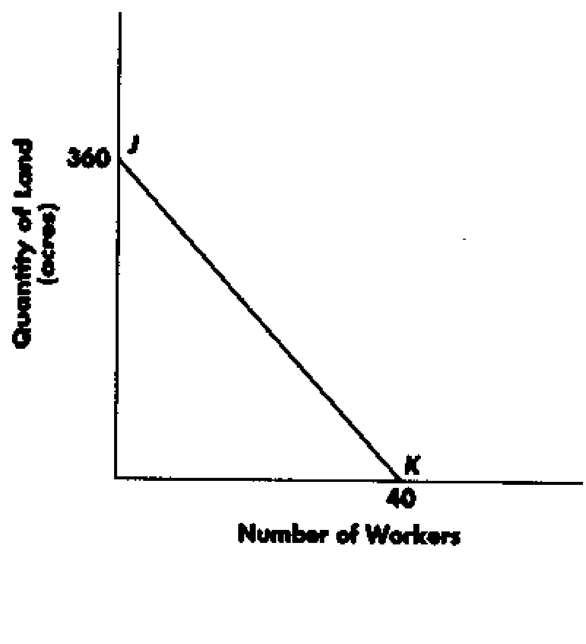
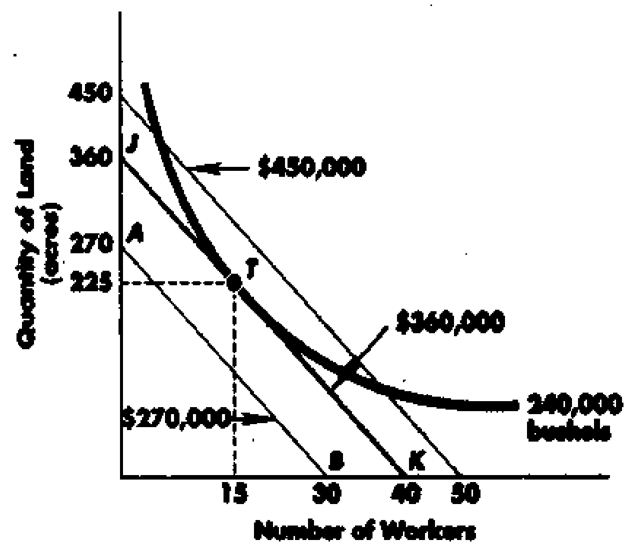


FIGURE 7-12 COST MINIMIZATION

The least costly way to produce 240,000 bushels of wheat is shown by point *T*, where the production indifference curve is tangent to the budget line, *JK*. Here the farmer is employing 15 workers and using 225 acres of land. It is not possible to produce 240,000 bushels on a smaller budget, and any larger budget would be wasteful.

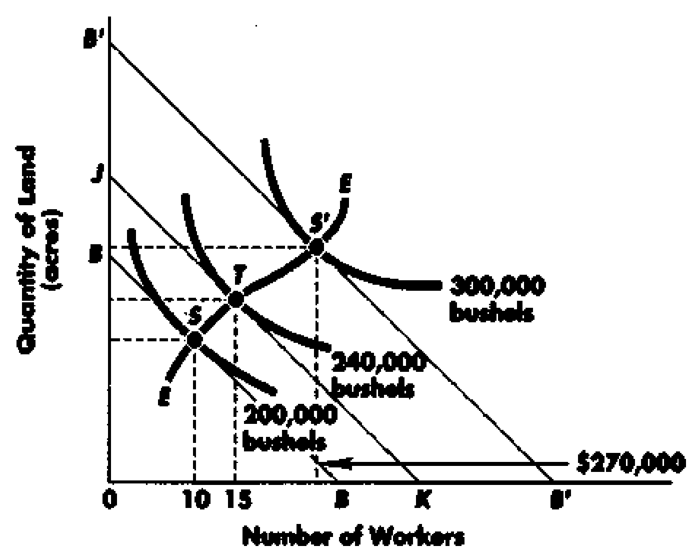


240,000 bushels of output. The farmer can repeat this procedure exactly for any other output quantity, such as 200,000 bushels or 300,000 bushels. In each case, we draw the corresponding production indifference curve and find the lowest budget line that permits the farm to produce that much. For example, in Figure 7-13, budget line *BB* is tangent to the indifference curve for 200,000 units of output; budget line *JK* is, again, tangent to the indifference curve for 240,000 bushels; and budget line *B'B'* is tangent to the indifference curve for 300,000 units of output. This gives us three tangency points: *S*, which gives the input combination that produces a 200,000-bushel output at lowest cost; *T*, which gives the same information for a 240,000-bushel output; and *S'*, which indicates the cost-minimizing input combination for the production of 300,000 bushels.

This process can be repeated for as many other levels of output as we like. For each such output we draw the corresponding production indifference curve and find its point of tangency with a

FIGURE 7-13 THE FIRM'S EXPANSION PATH

Each point of tangency, such as *S*, between a production indifference curve and a budget line shows the combination of inputs that can produce the output corresponding to that indifference curve at lowest cost. The locus of all such tangency points is *EE*, the firm's expansion path.



budget line. Orange curve *EE* in Figure 7-13 connects all these cost-minimizing points; that is, it is the locus of *S*, *T*, *S'*, and all other points of tangency between a production indifference curve and a budget line. Curve *EE* is the firm's expansion path.

The **expansion path** is the locus of the firm's cost-minimizing input combinations for all relevant output levels.

Point *T* in Figure 7-12 shows the quantity of output (given by the production indifference curve through that point) and the total cost (shown by the tangent budget line). Similarly, we can determine the output and total cost for every other point on the expansion path, *EE* in Figure 7-13. For example, at point *S*, output is 200,000 and total cost is $270,000. This is precisely the sort of information we need to find the firm's total cost curve; that is, it is just the sort of information contained in Table 7-4, the source of the total cost curve and the average and marginal cost curves in Figure 7-4. Thus:

The points of tangency between a firm's production indifference curves and its budget lines yield its expansion path, which shows the firm's cost-minimizing input combination for each pertinent output level. This information also yields the output and total cost for each point on the expansion path, which is just what we need to draw the firm's cost curves.

EFFECTS OF CHANGES IN INPUT PRICES

Suppose that the cost of renting land increases and the wage rate of labor decreases. This means that the budget lines will differ from those depicted in Figure 7-12. Specifically, with land now more expensive, any given sum of money will rent fewer acres, so the intercept of each budget line on the vertical (land) axis will shift *downward*. Conversely, with labor cheaper, any given sum of money will buy more labor, so the intercept of the budget line on the horizontal (labor) axis will shift to the *right*. Figure 7-14 depicts a series of budget lines corresponding to a $1,500 per acre rental rate for land and a $6,000 annual wage for labor. These budget lines are less steep than those shown in Figure 7-12, and point *E* now represents the least costly way to produce 240,000 bushels of wheat.

To assist you in seeing how things change, Figure 7-15 combines, in a single graph, budget line *JK* and tangency point *T* from Figure 7-12 with budget line *WV* and tangency point *E* from Figure 7-14. Notice that point *E* lies below and to the right of *T*, meaning that as wages decrease and rents increase, the firm will hire more labor and rent less land. As common sense suggests, when the price of one input rises in comparison with that of another, it will pay the firm to use less of the more expensive input and more of the other.

In addition to substituting one input for another, a change in the price of an input may induce the firm to alter its level of output. This is a subject that we will cover in the next chapter.

FIGURE 7-14 OPTIMAL INPUT CHOICE AT A DIFFERENT SET OF INPUT PRICES

If input prices change, the combination of inputs that minimizes costs will normally change, too. In this diagram, land rents for $1,500 per acre (more than in Figure 7-12) while labor costs $6,000 per year (less than in Figure 7-12). As a result, the least costly way to produce 240,000 bushels of wheat shifts from point *T* in Figure 7-12 to point *E* here.

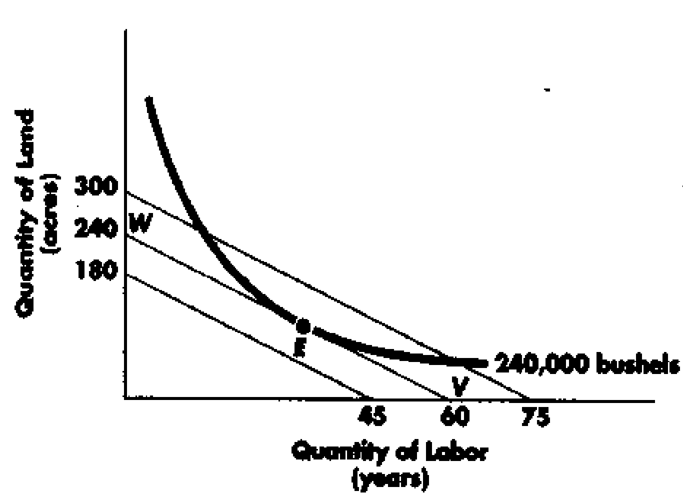
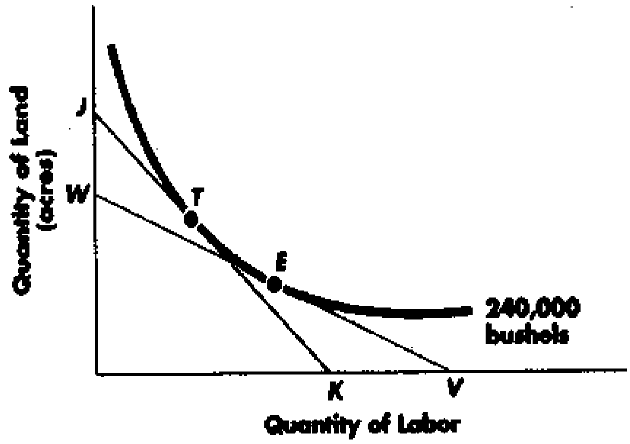


FIGURE 7-15 HOW CHANGES IN INPUT PRICES AFFECT INPUT PROPORTIONS

When land becomes more expensive and labor becomes cheaper, the budget lines (such as *JK*) become less steep than they were previously (see *WV*). As a result, the least costly way to produce 240,000 bushels shifts from point *T* to point *E*, at which the firm uses more labor and less land.



SUMMARY

1. A production function can be illustrated by a series of **production indifference curves,** each of which shows all the input combinations capable of producing a specified amount of output.
2. As long as each input has a positive marginal physical product, production indifference curves will have negative slopes and the higher curves will represent larger amounts of output than the lower curves. Because of diminishing returns, these curves characteristically bend toward the origin near the middle.
3. The optimal input combination for any given level of output is indicated by the point of tangency between a budget line and the corresponding production indifference curve.
4. The firm's **expansion path** shows, for each of its possible output levels, the combination of input quantities that minimizes the cost of producing that output.
5. Total cost for each output level can be derived from the production indifference curves and the budget lines tangent to them along the expansion path. These figures can be used to determine the firm's total cost, average cost, and marginal cost curves.
6. When input prices change, firms will normally use more of the input that becomes relatively less expensive and less of the input that becomes relatively more expensive.

KEY TERMS

Production indifference curve   Budget line   Expansion path

QUESTIONS FOR REVIEW

1. Compound Consolidated Corporation (CCC) produces containers using two inputs: labor and glue. If labor costs \$5 per hour and glue costs \$5 per gallon, draw CCC's budget line for a total expenditure of \$100,000. In this same diagram, sketch a production indifference curve indicating that CCC can produce no more than 1,000 containers with this expenditure.
2. With respect to Review Question 1, suppose that wages rise to \$10 per hour and glue prices rise to \$6 per gallon. How are CCC's optimal input proportions likely to change? (Use a diagram to explain your answer.)
3. What happens to the expansion path of the firm in Review Question 2?

CHAPTER 8

OUTPUT, PRICE, AND PROFIT: THE IMPORTANCE OF MARGINAL ANALYSIS

Business is a good game. . . . You keep score with money.
Nolan Bushnell, founder of Atari, one of the original video game manufacturers

We now approach the culmination of our discussion of the central building blocks of microeconomics: the analysis of the market mechanism. Chapters 5 and 6 dealt with the behavior of the first of the two main protagonists: the consumer. Then Chapter 7 introduced the other leading member of the cast: the firm. The firm's two main roles are, first, to produce its product efficiently and, second, to sell it at a profit. The previous chapter described production decisions. Now we turn to the selling decisions. How much should the firm sell and at what price, *given the constraining influence of consumer demand?*

An **OPTIMAL DECISION** is one which, among all the decisions that are actually possible, is best for the decision maker. For example, if profit is the sole objective of some firm, the price that makes the firm's profit as large as possible is optimal for that company.

These four chapters of Part II all describe how firms and consumers can make decisions that are **optimal,** meaning that they go as far as possible, given the circumstances, to promote the goals that the decision maker happens to have selected. In this chapter, we will continue to assume that the goal of the business firm is maximization of its total profit, just as the objective of the consumer was assumed to be maximization of his or her utility. As in the previous three chapters, marginal analysis is the tool that helps us to determine what constitutes an optimal decision. This chapter summarizes and generalizes what we have learned about the methods of marginal analysis, showing also how they can be used to deal with other issues in which optimality is a key issue.

The analysis leads to some surprising conclusions which show that unaided common sense can sometimes be misleading in business decisions. For example, suppose that a firm suffers a sharp increase in its rent or some other fixed cost. How should the firm react? Some would argue that the firm should raise the price of its product to cover the higher rent, but others would argue that it should cut its price in order to increase its sales enough to pay the higher rent. We will see in this chapter that *both* of these answers are incorrect! A profit-maximizing firm faced with an increase in its rent should neither raise nor lower its price.

TWO ILLUSTRATIVE CASES

Price and output decisions can perplex even the most experienced business people, as the following real-life illustrations show. At the end of the chapter, we will see how the tools described here helped solve the problems.

CASE 1: MAKING PROFITS BY SELLING BELOW COSTS

Unaided common sense can sometimes be misleading in business decisions. Our first illustrative case indicates that it is possible for a firm to make a profit by selling at a price that is apparently below its cost!

In a recent legal battle between two manufacturers of pocket calculators, Company B accused Company A of selling 10 million sophisticated calculators at a price of $12, which A allegedly knew was too low to cover its costs. B claimed that A was doing this only to drive B out of business. At first, Company A's records, which were revealed to the court, appeared to confirm B's accusations. The cost of materials, labor, advertising, and other direct costs of the calculator came to $10.30 per calculator. Company A's accountants also assigned to this product its share of the company's annual expenditure on overhead—such items as general administration, research, and the like—which amounted to $4.25 per calculator. The $12 price clearly did not cover the $14.55 cost attributed to each calculator. Yet economists representing Company A were able to convince the court that manufacturing the calculator was a profitable activity for Company A, so that there was no basis on which to conclude that its only purpose was to destroy B. At the end of the chapter, we will explain just how this was possible.

CASE 2: PRICING A SIX-PACK

The managers of one of America's largest manufacturers of soft drinks became concerned when a rival company introduced a cheaper substitute for one of their leading products. As a result, some of the market leader's managers advocated a reduction in the price of a six-pack from $1.50 to $1.35. This stimulated a heated debate. It was agreed that the price should be cut if it was not likely to reduce the company's profits. Although some of the managers maintained that the cut made sense because of the additional sales it would stimulate, others held that the price cut would hurt the company by cutting profit per unit of output. The company had reliable information about costs, but knew rather little about the shape of its demand curve. At this point a group of consultants, including one of the authors of this book, was called in to offer their suggestions. We will see how economic analysis enabled them to solve the problem even though the vital demand elasticity figures were unavailable.

PRICE AND QUANTITY: ONE DECISION, NOT TWO

When a firm such as Intel introduces a new line of computer chips, it must decide on the prices it will charge and the number of chips it will produce. These are clearly crucial decisions. They have a vital influence on the firm's labor requirements, on consumer response to the product, and, indeed, on the future success of the company. This chapter describes tools that a firm like Intel can use to choose the price-output combination that makes its operations as profitable as they can be.

When firms select a *price* and a *quantity* of output that maximize profits, it seems that they must choose two numbers. In fact, however, they can pick only one. Once they have selected the *price*, the *quantity* they will sell is up to consumers. Alternatively, firms may decide *how much* they want to sell, but then they must leave it to the market to determine the *price* at which this quantity can be sold. These observations are important because they illustrate explicitly the powerful role that consumers play in the market. Management gets two numbers by making only one decision because the firm's demand curve tells it, for any quantity it may decide to market, the highest possible price its product can bring.

To illustrate, we return once more to our hypothetical firm, Florence's poultry farm. Florence, who sells mostly to consumers in the neighborhood, is exploring what she can do to make a better living from the enterprise. Whether or not Flo is aware of it, there is a demand curve for her chickens, *DD* in Figure 8-1. This curve shows that if she becomes greedy and decides to charge the relatively high price of $19 for a 20-pound package of chicken (point *a* on the curve), she can sell only 1 package per week. On the other hand, if she wants to sell as many as 5.5 packages per week, she can find the required number of customers only by offering chicken at the lower price of $14.50 (point *k*). In summary:

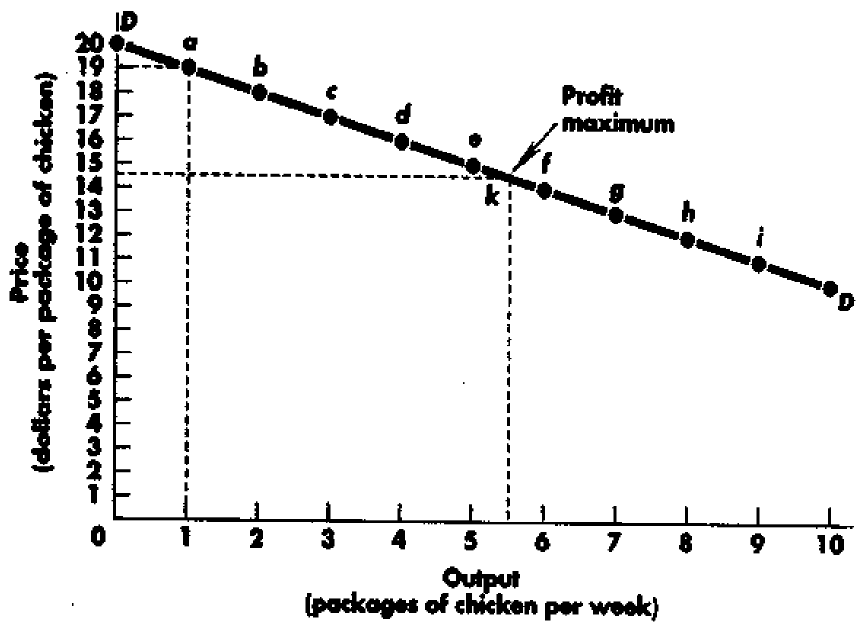
Each point on the demand curve represents a price-quantity pair. The firm can pick any such pair. But it can never pick the price corresponding to one point on the demand curve and the quantity corresponding to another point, since such an output could not be sold at the selected price.

Throughout this chapter, then, we will not discuss price and output decisions separately, for they are merely two different aspects of the same decision. To analyze this decision, we will make a strong assumption about the behavior of business firms—the assumption that firms strive for the largest possible total profit

FIGURE 8-1

This graph shows the quantity of product demanded at each price. For example, the curve shows that at a price of $14.50 (point *k*), 5.5 units will be demanded.

FLO FARMER'S DEMAND CURVE



to the exclusion of any other goal. While not literally correct, this seems to be a useful simplification of a much more complex reality. At least economists have long believed so.

DO FIRMS REALLY MAXIMIZE PROFITS?

Naturally, many people have questioned whether firms really try to maximize profits to the exclusion of all other goals. Japanese companies, for instance, are well-known for their other corporate priorities such as life-time employment for workers and a preference for doing business with other Japanese companies.

Further, business people are like other human beings: Their motives are varied and complex. Given the choice, many executives might prefer to control the largest firm rather than the most profitable one. Some may be fascinated by technology and therefore may spend so much on R&D that it cuts down on profit. Others may be motivated by a desire to "do good" and therefore give away some of the stockholders' money to hospitals and colleges. Different managers within the same firm may not always agree with one another on goals, so that it may not even make sense to speak about "*the* goal of the firm." Thus, any attempt to summarize the objectives of management in terms of a single number (profit) is bound to be an oversimplification.

In addition, the exacting requirements for maximizing profits are tough to satisfy. In deciding how much to invest, what price to set for a product, or how much to allocate to the advertising budget, the range of available alternatives is enormous. Also, information about each alternative is often expensive and difficult to acquire. As a result, when a firm's management decides on an $18 million construction budget, it rarely compares the consequences of that decision in any detail with the consequences of all the possible alternatives—such as budgets of $17 million or $19 million. But unless *all* the available possibilities are compared, there is no way management can be sure that it has chosen the one that brings in the highest possible profit.

Often, management's concern is *whether the decision's results are likely to be acceptable*—whether its risks will be acceptably low, whether its profits will be acceptably high—so that the company can live satisfactorily with the outcome. Such analysis cannot be expected to bring in the maximum possible profit. The decision may be good, but some unexplored alternative may be even better.

Decision making that seeks only solutions that are *acceptable* has been called *satisficing*, to contrast it with optimizing (profit maximization). Some analysts, such as Nobel prize winner Herbert Simon of Carnegie-Mellon University, have concluded that decision making in industry and government is often of the satisficing variety.

But even if this is true, it does not necessarily make profit maximization a bad assumption. Recall our discussion of abstraction and model-building in Chapter 1. A map of Los Angeles that omits hundreds of roads is no doubt "wrong" if interpreted as a literal description of the city. Nonetheless, by capturing the most important elements of reality, it may help us understand the city better than a map that is cluttered with too much detail. Similarly, we can learn much about the behavior of business firms by assuming that they try to maximize profits, even though we know that not *all* of them act this way *all* of the time.

We will therefore assume throughout this chapter and for most of the book that the firm has only one objective. It wants to make its *total* profit as large as possible. Our analytic strategy will be to determine what output level (or price)

achieves this goal. But you should keep in mind that many of the results depend on a simplifying assumption, so the conclusions will not apply to every case. Our decision to base the analysis on the assumption that the firm maximizes profits gives us sharper insights, but we pay with some loss of realism.

TOTAL PROFIT: KEEP YOUR EYE ON THE GOAL

The **TOTAL PROFIT** of a firm is its net earnings during some period of time. It is equal to the total amount of money the firm gets from sales of its products (the firm's total revenue) minus the total amount that it spends to make those products (total cost).

Total profit, then, is assumed to be the goal of the firm. It is, by definition, the difference between what the company earns in the form of sales revenue and what it pays out in the form of costs:

$$\text{Total profit} = \text{Total revenue} - \text{Total costs}$$

Total profit defined in this way is called *economic profit* to distinguish it from the accountant's definition of profit. The two concepts of profit differ because total cost, in the economist's definition, includes the opportunity cost of any capital, labor, or other inputs supplied by the owner of the firm. Thus, if a small business earns just enough to pay the owner the money that her labor and capital could have earned if they had been sold to others (say, $60,000 per year), economists say that she is earning zero *economic* profit. (She is just covering *all* her costs, including her opportunity costs.) In contrast, most accountants will say her profit is $60,000.

To analyze how total profit depends on output, we must therefore study the behavior of the two components of total profit: total revenue (TR) and total cost (TC). It should be obvious that both total revenue and total cost depend on the output-price combination the firm selects; it is these relationships that we study next.

TOTAL, AVERAGE, AND MARGINAL REVENUE

Total revenue can be calculated directly from the demand curve since, by definition, it is the product of price times the quantity that will be bought at that price:

$$\text{TR} = P \times Q$$

Table 8-1 shows how the total revenue schedule is derived from the demand schedule for Flo's farm. The first two columns simply express the demand curve of Figure 8-1 in tabular form. The third column gives, for each quantity, the product of price times quantity. For example, if Flo sells 6 packages of chicken at a price of $14 per pack, her weekly sales revenue will be 6 packages × $14 per package = $84.

Figure 8-2 displays Flo's total revenue schedule in graphical form as the black TR curve. This graph shows precisely the same information as the demand curve in Figure 8-1, but in a somewhat different form. For example, point *e* on the demand curve in Figure 8-1, which shows a price-quantity combination of P = $15 per package and Q = 5 packages of chicken per week, appears as point *E* in Figure 8-2 as a total revenue of $75 per week ($15 per package times 5 packages). Similarly, each point on the TR curve in Figure 8-2 corresponds to the similarly labeled point in Figure 8-1.

The relationship between the demand curve and the TR curve can be rephrased in a slightly different way. Since the price of the product is the revenue

The **AVERAGE REVENUE (AR)** is total revenue (TR) divided by quantity.

per unit that the firm receives, we can view the demand curve as the curve of average revenue (AR). Average revenue and total revenue (TR) are related to one another in a simple way.[1] Specifically:

$$AR = TR/Q = P \times Q/Q = P$$

TABLE 8-1

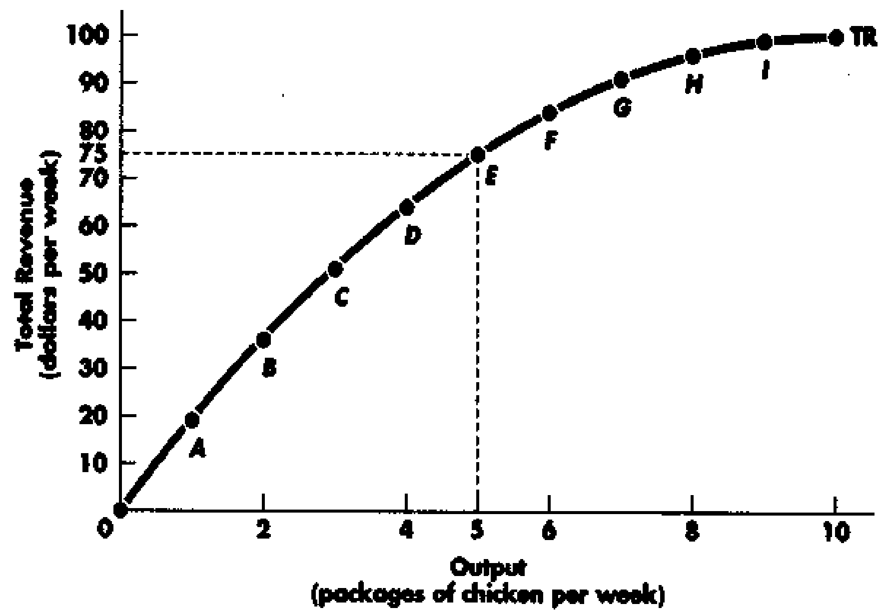
FLO FARMER'S DEMAND SCHEDULE, TOTAL REVENUE SCHEDULE, AND MARGINAL REVENUE SCHEDULE

Number of Chicken Packages (per week)	Price = Average Revenue ($ per package)	Total Revenue ($ per week)	Marginal Revenue ($ per package)
0	—	0	
			19
1	19	19	
			17
2	18	36	
			15
3	17	51	
			13
4	16	64	
			11
5	15	75	
			9
6	14	84	
			7
7	13	91	
			5
8	12	96	
			3
9	11	99	
			1
10	10	100	

Note: Data correspond to Figure 8-1.

FIGURE 8-2

FLO'S TOTAL REVENUE CURVE



The total revenue curve for Flo's farm is derived directly from the demand curve, since total revenue is the product of price times quantity. Points A, B, C, D, E, and F in this diagram correspond to points a, b, c, d, e, and f, respectively, in Figure 8-1.

[1]See the appendix to this chapter for a general discussion of the relationship between totals and averages.

Therefore, average revenue and price are two names for the same thing. The reason should be clear. If a supermarket sells candy bars to any customer who wants them at the same price, say, $1, then the *average* revenue that the store derives from the sale of all these candy bars must also be $1.

MARGINAL REVENUE, often abbreviated MR, is the *addition* to total revenue resulting from the addition of 1 unit to total output. Geometrically, marginal revenue is the *slope* of the total revenue curve. Its formula is $MR_1 = TR_1 - TR_0$, and so on.

Finally, the last column of Table 8-1 shows what is called the **marginal revenue (MR)** for each level of output—an analytic tool whose use will be explained presently. This concept (analogous to marginal utility and marginal cost) refers to the *addition* to total revenue that results from raising output by 1 unit. Thus, in Table 8-1, we see that when output rises from 2 to 3 units, total revenue goes up from $36 to $51, so that marginal revenue is $51 minus $36, or $15.

TOTAL, AVERAGE, AND MARGINAL COST

The revenue side is, of course, only half of the firm's profit picture. We must turn to the cost side for the other half. As we saw in the last chapter, average cost (AC) and marginal cost (MC) are obtained directly from total cost (TC) in exactly the same way that average and marginal revenue were calculated from total revenue.

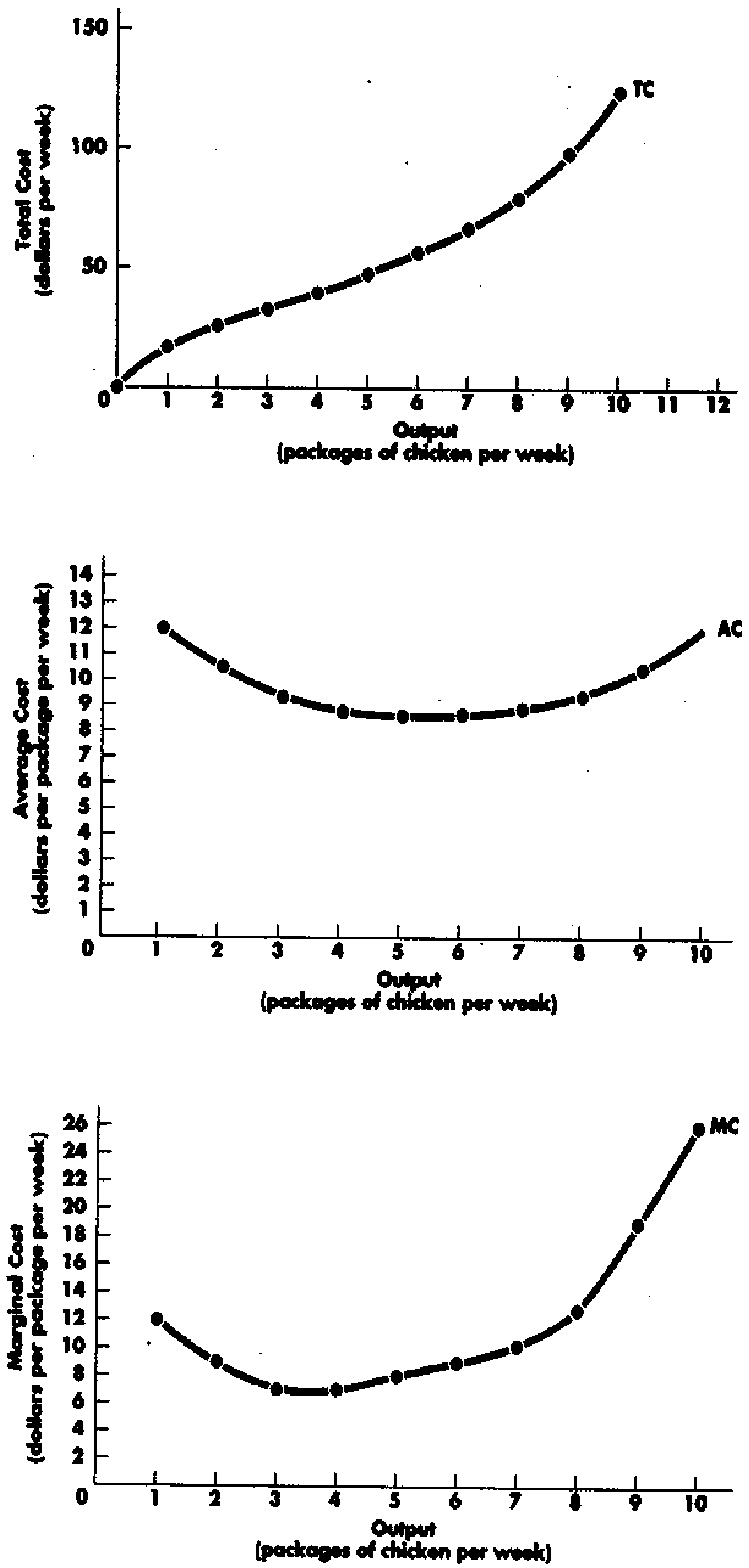
Figure 8-3 on the next page plots the numbers in Table 8-2 and thus shows the total, average, and marginal cost curves for the poultry farm. As we learned in the last chapter, the shapes of the curves depicted here are considered typical. The shapes mean that, in any given industry, there is some size of firm that is most efficient in producing the output. Smaller enterprises lose any advantages that derive from a large volume of production, and so their average cost (the cost per unit of output) will be greater than that of a firm operating at the most efficient size of output. Similarly, firms that are too large will suffer from difficulties of supervision and coordination, and perhaps from bureaucratic controls, so that their costs per unit of output will also be higher than those of a firm of the most efficient size.

TABLE 8-2 FLO'S TOTAL, AVERAGE, AND MARGINAL COSTS

Number of Chicken Packages (per week)	Total Cost ($ per week)	Marginal Cost ($ per package)	Average Cost ($ per week)
0	0.0		—
1	17.0	17.0	17.0
2	26.0	9.0	13.0
3	33.0	7.0	11.0
4	40.0	7.0	10.0
5	48.0	8.0	9.6
6	57.0	9.0	9.5
7	67.2	10.2	9.6
8	80.0	12.8	10.0
9	99.0	19.0	11.0
10	125.0	26.0	12.5

FIGURE 8-3 FLO'S COST CURVES

The graphs show, for each possible level of output, Flo's total cost, average cost, and marginal cost. These graphs are reproduced from Figure 7-4 in the previous chapter.



MAXIMIZATION OF TOTAL PROFIT

We now have all the tools we need to answer our central question: What combination of output and price will yield the largest possible total profit? To study how total profit depends on output, we bring together in Table 8-3 the total revenue and total cost schedules from the two previous tables. The last column in Table 8-3—called, appropriately enough, total profit—is just the difference between total revenue and total cost at each level of output.

Since we assume that Flo's objective is to maximize profits, it is simple enough to determine the level of production she will choose. By producing and selling between 5 and 6 packages of meat per week, the farm earns the highest level of profits it is capable of earning—some $27 per week. Any higher or lower rate of production would lead to lower profits. For example, weekly profits would drop to $16 if output were increased to 8 packages. If Flo were to make the mistake of producing 10 packages per week, she would actually have a net weekly loss of $25.

PROFIT MAXIMIZATION: A GRAPHICAL INTERPRETATION

We can present the same information on a graph. In Panel (a) of Figure 8-4, we bring together into a single diagram the total revenue curve from Figure 8-2 and the total cost curve from Figure 8-3. Total profit, which is the difference between total revenue and total cost, appears in the diagram as the *vertical* distance between the TR and TC curves. For example, when output is 4 units, total revenue is $64 (point *A*), total cost is $40 (point *B*), and total profit is the distance between points *A* and *B*, or $24 per week.

In this graphical view of the problem, Flo wants to maximize total profit, which is the vertical distance between the TR and TC curves. Panel (b) of Figure 8-4 shows the curve of total profit—that is, TR minus TC. We see that it reaches its maximum value, about $27, at an output level between 5 and 6

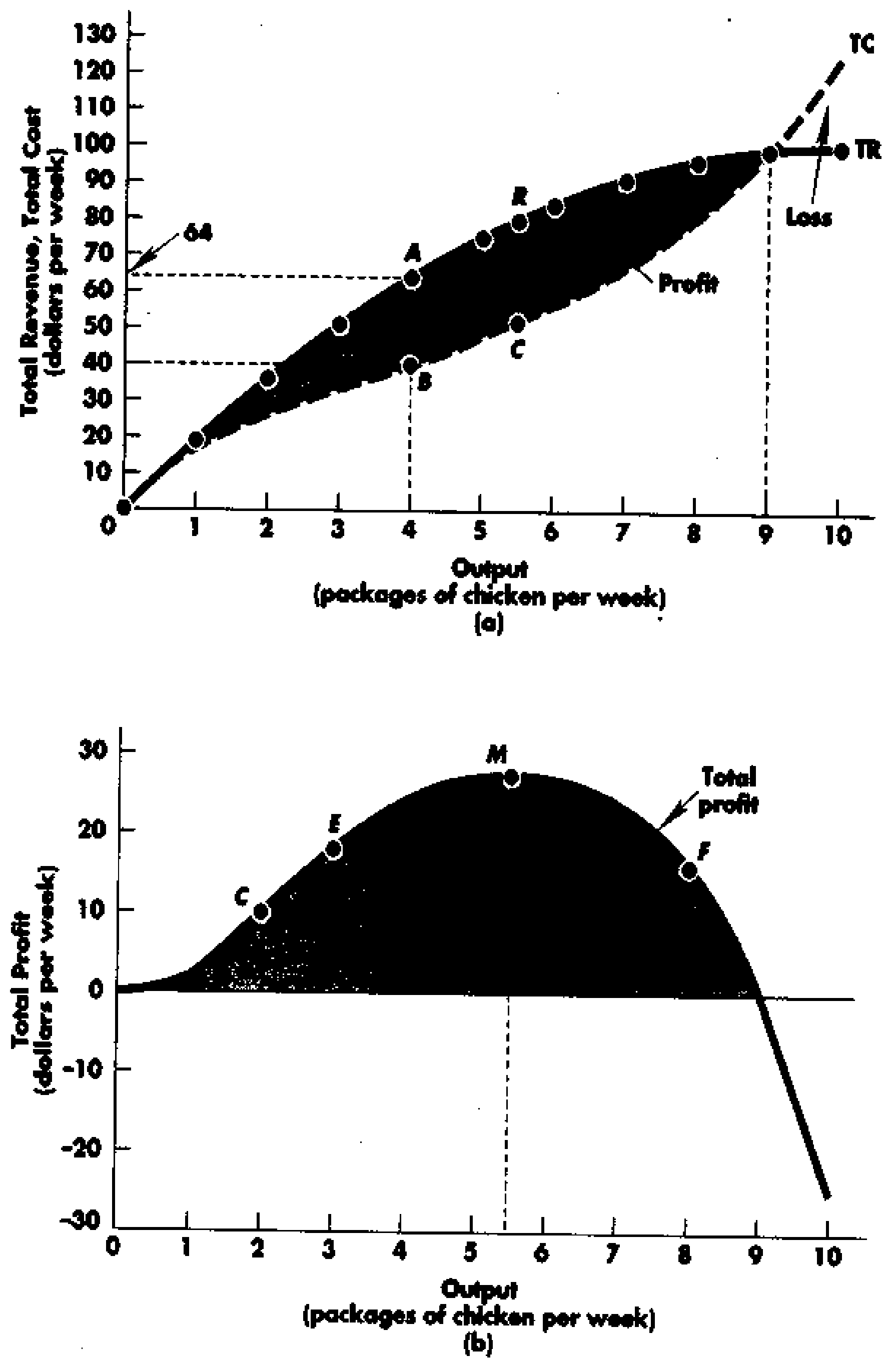
TABLE 8-3

FLO'S TOTAL REVENUES, COSTS, AND PROFIT

Chicken Packages (per week)	Total Revenue (TR)	Total Cost (TC) (per week)	Total Profit (TP)
0	0	$ 0	0
1	19	17.0	2.0
2	36	26.0	10.0
3	51	33.0	18.0
4	64	40.0	24.0
5	**75**	**48.0**	**27.0**
6	**84**	**57.0**	**27.0**
7	91	67.2	23.8
8	96	80.0	16.0
9	99	99.0	0
10	100	125.0	-25.0

FIGURE 8-4 PROFIT MAXIMIZATION: A GRAPHICAL INTERPRETATION

Flo's profits are maximized when the vertical distance between her total revenue curve, TR, and total cost curve, TC, is at its maximum. In Panel (a) this occurs at an output of 5.5 units per week; total profits are CR, or $27. The total profit curve is shown in Panel (b). Naturally, it reaches its maximum value (about $27) at 5.5 units (point *M*).



units per week. This is the same conclusion we reached by looking at Table 8-3.

The total profit curve in Figure 8-4(b) is shaped like a hill. Though such a shape is not inevitable, we expect a hill shape to be typical for the following reason. If a firm produces nothing, it certainly earns no profit. At the other extreme, a firm can produce so much output that it swamps the market, forcing price down so low that it loses money. Only at intermediate levels of output—something between zero and the amount that floods the market—will the company

earn a positive profit. Consequently, the total profit curve will rise from zero (or negative) levels at a very small output to positive levels at intermediate outputs; finally, it will fall to negative levels when output gets too large.

MARGINAL ANALYSIS AND MAXIMIZATION OF TOTAL PROFIT

We see from Figure 8-4 and Table 8-3 that there may be many levels of output that yield a positive profit, but the firm is not aiming for just *any* level of profit. It wants the *largest* possible profit. If management knew the exact shape of its profit hill, that is, if it had Table 8-3, choosing the optimal level of output would be a simple task indeed. It would only have to locate the point, such as *M* in Figure 8-4(b), that defined the top of its profit hill. However, management rarely if ever has so much information, so a different technique for finding the optimum is required. That technique is marginal analysis—the same set of tools we used to analyze the firm's input purchase decisions in Chapter 7 and the consumer's buying decisions in Chapters 5 and 6.

MARGINAL PROFIT is the *addition* to total profit resulting from one more unit of output.

This time we will use a concept known as **marginal profit** to solve Flo's problem. Referring back to Table 8-3 on page 189, we see that an increase in Flo's weekly output from 2 to 3 packages of chicken would raise total profit from $10 to $18; that is, it would generate $8 in *additional* profit. We call this the *marginal profit* resulting from the addition of the third unit. Similarly, marginal profit from the seventh unit would be:

Total profit from 7 units − Total profit from 6 units = $23.80 − $27.00 = −$3.20

The marginal rule for finding the optimal level of output is easy to understand:

If the marginal profit from increasing output by 1 unit is positive, then output should be increased. If the marginal profit from increasing output by 1 unit is negative, then output should be decreased. Thus, an output level can maximize total profit only if marginal profit equals zero at that output.

On Flo's Farm, the marginal profit from the third unit is $8. This means that going from the second to the third unit *adds* $8 to profit, so it pays to produce the third unit. But marginal profit from the seventh unit is $23.80 − $27.00, or −$3.20, so the firm should not produce the seventh package because that would reduce total profit by $3.20. Only where marginal profit is neither positive nor negative (as is true for the sixth unit of output) can total profit be as big as possible, because neither increasing nor reducing output can add to total profit.

The profit hill in Figure 8-4(b) is a graphical representation of the condition that marginal profit should equal zero. Marginal profit is defined as the additional profit that accrues to the firm when output rises by 1 unit. So, when output is increased, say, from 2 units to 3 units, the distance *CD* in Figure 8-4(b), total profit rises by $8.00 (the distance *DE*) and marginal profit is therefore *DE*/*CD*. This is precisely the definition of the *slope* of the total profit curve between points *C* and *E*. In general:

Marginal profit is the slope of the total profit curve.

With this geometric interpretation in hand, we can easily understand the logic of the marginal profit rule. At a point such as *C*, where the total profit curve is

rising, marginal profit (which equals slope) is positive. Profit cannot be maximal at such a point, because we can increase profits by moving farther to the right. A firm that decided to stick to point C would be wasting the opportunity to increase profits by increasing output. Similarly, the firm cannot be maximizing profits at a point like F, where the slope of the curve is negative, because there marginal profit (which, again, equals slope) is negative. If it finds itself at a point like F, the firm can raise its profit by decreasing its output.

Only at a point such as M, where the total profit curve is neither rising nor falling, can the firm possibly be at the top of the profit hill rather than on one of the sides of the hill. Point M is precisely where the slope of the curve—and hence the marginal profit—is zero. Thus:

An output decision cannot be optimal unless the corresponding marginal profit is zero.

The firm is not interested in marginal profit for its own sake, but rather for what it implies about *total* profit. Marginal profit is like the needle on the temperature gauge of a car: the needle itself is of no concern to anyone, but failure to watch it can result in serious consequences.

One common misunderstanding of marginal analysis is the idea that it seems foolish to go to a point where marginal profit is zero. "Isn't it better to earn a positive marginal profit?" This notion springs from a confusion between the quantity one is seeking to maximize (*total* profit) and the gauge that indicates whether such a maximum has in fact been attained (*marginal* profit). Of course, it is better to have a positive *total* profit than a zero total profit. Still, a zero value on the *marginal* profit gauge merely indicates that all is well, that *total* profit is at its maximum.

MARGINAL REVENUE AND MARGINAL COST: GUIDES TO AN OPTIMUM

An alternative version of the marginal analysis of profit maximization can be derived from the cost and revenue components of profit. For this purpose, refer back to Figure 8-4 on page 190, where we used total revenue (TR) and total cost (TC) curves to construct the profit hill. Note that there is another way of finding the profit-maximizing solution.

We want to maximize the vertical distance between the TR and TC curves. This distance, we see, is not maximal at an output level such as 2 units, because there the two curves are growing farther apart. If we move farther to the right, the vertical distance between them (which is total profit) will increase. Conversely, we have not maximized the vertical distance between TR and TC at an output level such as 8 units, because there the two curves are coming closer together. We can add to profit by moving farther to the left (reducing output). The conclusion from the graph, then, is that total profit—the vertical distance between TR and TC—is maximized only when the two curves are neither growing farther apart nor coming closer together, that is, when their *slopes* are equal.

Marginal revenue and marginal cost curves, which we learned about previously (see pages 186–187 and 187–189), will help us understand this concept better. For precisely the same reason that marginal profit is the slope of the total profit curve, marginal revenue is the slope of the total revenue curve, since it represents the increase in total revenue resulting from the sale of 1 additional unit. Again, for the same reason, marginal cost is equal to the slope of the total

TABLE 8-4

FLO'S MARGINAL REVENUE AND MARGINAL COST

Chicken Packages (per week)	Marginal Revenue (per package per week)	Marginal Cost (per package per week)
0		
	$19	$17.0
1		
	17	9.0
2		
	15	7.0
3		
	13	7.0
4		
	11	8.0
5		
	9	**9.0**
6		
	7	10.2
7		
	5	12.8
8		
	3	19.0
9		
	1	26.0
10		

cost curve. This interpretation of marginal revenue and marginal cost, respectively, as the *slopes* of the total revenue and total cost curves permits us to restate the geometric conclusion we have just reached in an economically significant way:

Profit can be maximized only at an output level at which marginal revenue is (approximately) equal to marginal cost. In symbols:

$$MR = MC$$

The logic of the MR = MC rule for profit maximization is straightforward.[2] When MR is *not* equal to MC, profits cannot possibly be maximized because the firm can increase its profits either by raising its output or by reducing it. For example, if MR = \$15 and MC = \$7, an additional unit of output *adds* \$15 to revenues but only \$7 to costs. Hence, the firm can increase its net profit by \$8 by producing and selling one more unit. Similarly, if MC exceeds MR, say, MR = \$3 and MC = \$19, then the firm loses \$16 on its marginal unit, so it can add \$16 to its profit by reducing output by 1 unit. Only when MR = MC is it impossible for the firm to add to its profit by changing its output level.

Table 8-4 reproduces marginal revenue and marginal cost data for Flo's farm from Tables 8-1 and 8-2. The table shows, as must be true, that the MR = MC rule leads us to the same conclusion as Figure 8-4 and Table 8-3. If she wants to maximize her profits, Flo should produce and sell between 5 and 6 packages of chicken per week. The marginal revenue of the fifth package of chicken is \$11 (\$75 from the sale of 5 packages less \$64 from selling 4) while the marginal cost is only \$8 (\$48 − \$40). Therefore, the firm should produce the fifth unit. The seventh package, however, brings in only \$7 in marginal revenue while its marginal cost is \$10.20—clearly a losing proposition. Only between 5 and 6 units of output does MR = MC = \$9.

Because the graphs of marginal analysis will prove so useful in later chapters, Figure 8-5(a) shows the MR = MC condition for profit maximization graphically. The black curve labeled MR in the figure is the marginal revenue schedule from Table 8-4. The orange curve labeled MC is the marginal cost schedule. They intersect at point *E*, which is therefore the point where marginal revenue and marginal cost are equal. The optimal output for Flo is between 5 and 6 units.[3] Figures 8-5(b) and 8-5(c), respectively, are reproductions of the TR and TC curves from the upper panel of Figure 8-4 and the total profit curve from the lower panel of that figure. Note how MC and MR intersect at the same output at which the distance of TR above TC is greatest, which is the output at which the profit hill reaches its peak.

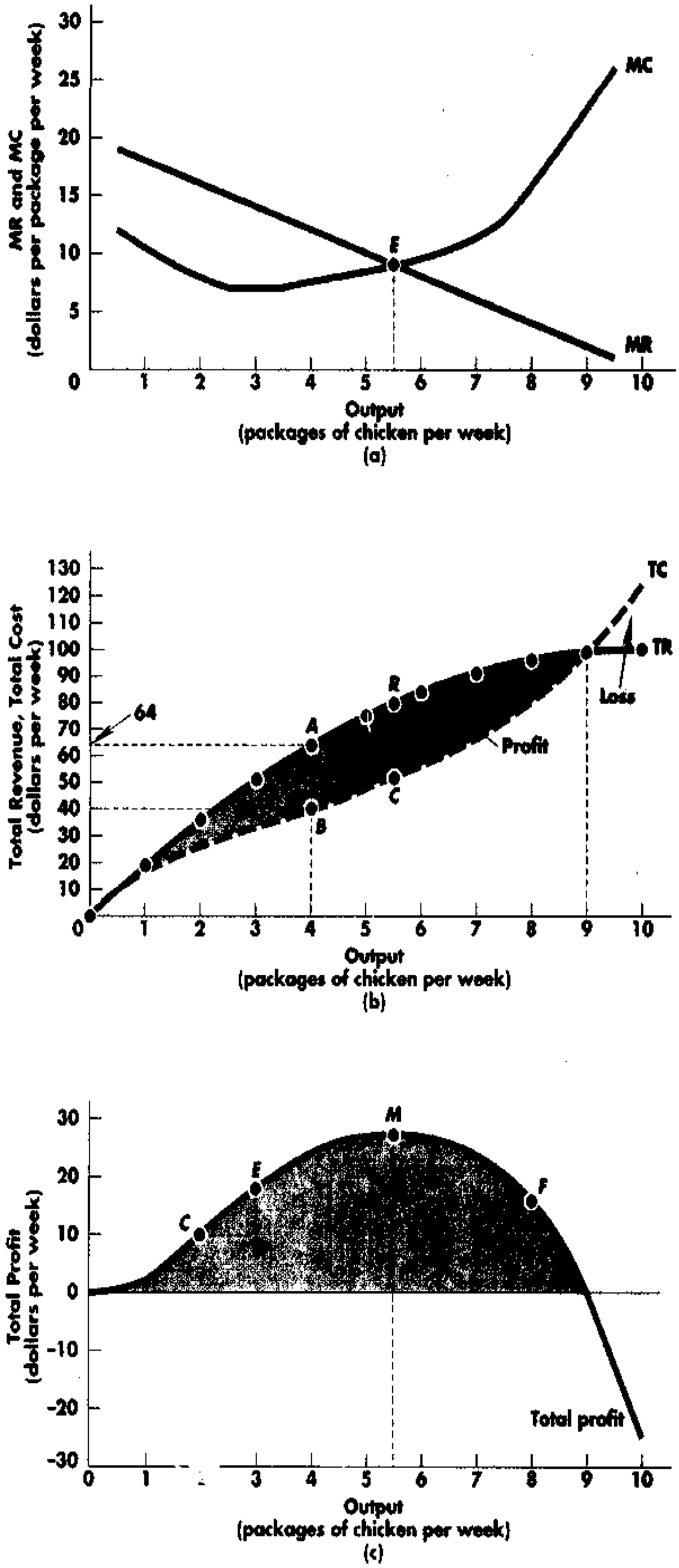
---

[2]You may have surmised by now that just as Total profit = Total revenue − Total cost, it must be true that Marginal profit = Marginal revenue − Marginal cost. This is, in fact, correct. It also shows that when Marginal profit = 0, we must have MR = MC.

[3]We must note one important qualification. Sometimes marginal revenue and marginal cost curves do not have the nice shapes depicted in Figure 8-5(a), and they may intersect more than once. In such cases, while it remains true that MR = MC at the output level that maximizes profits, there may be other output levels at which MR is also equal to MC but at which profits are not maximized.

FIGURE 8-5 PROFIT MAXIMIZATION: ANOTHER GRAPHICAL INTERPRETATION

Profits are maximized where marginal revenue (MR) is (approximately) equal to marginal cost (MC), for only at such a point will *marginal profit* be zero. Panel (a) shows the MR = MC condition for profit maximization graphically as point *E*, where output is 5.5 packages of chicken per week. The diagram also reproduces from Figure 8-4 the TR and TC curves in Panel (b) and the total profit curve in Panel (c), showing how all three agree that the profit-maximizing output is between 5 and 6 units



POLICY DEBATE PROFIT AND THE NEW MARKET ECONOMIES

The collapse of communism after its catastrophic failure to produce economic abundance has led the nations of eastern Europe and elsewhere to turn to the market mechanism. They hope that the market will soon bring them the sort of prosperity achieved by the industrialized countries.

The market, as we know, is driven by the profit motive, and in a free market, profits are not determined by a government agency, but by demand and cost conditions, as described by the demand and cost curves. Many citizens of these new market economies are appalled by the sizes of the profits that the free market affords to successful business people, and they are upset by the greed that these entrepreneurs display. There are pressures to put limits on these profits.

The same thing has happened in Great Britain and elsewhere as firms formerly owned by the government have been sold to private individuals and returned to the market. In Britain, a number of the privatized firms were initially monopolies, and the government has chosen to protect consumers by putting ceilings on prices *but not on profits*, in order to provide the firms with appropriate incentives. Yet, when some of these firms proved to be quite profitable, the British government agencies reduced the price ceilings in order to cut those profits, a move that was attacked sharply not only by the firms themselves, but also by some British economists. The debate in the United Kingdom and elsewhere amounts to this: should severe limits be placed on profits as a matter of fairness and to improve the ethical climate of society, or should this be avoided because ceilings on profits undermine the incentives for business success and therefore prevent the market mechanism from delivering the economic abundance of which it is capable?



WE HAVE DETERMINED PRICE AS WELL AS OUTPUT

At the start of this chapter, we set two goals: to determine the profit-maximizing *output* and also the profit-maximizing *price*. So far, we have found the profit-maximizing output, the output level at which MR = MC (5.5 packages of chicken per week in our example). That leaves us with the task of determining the profit-maximizing price.

Fortunately, this requires only one more easy step. As we said at the beginning of the chapter, once the firm has selected the output it wants to produce and sell, the demand curve determines the price it must charge to induce consumers to buy that amount of product. Consequently, if we know that the profit-maximizing output is 5.5 packages of chicken, the demand curve in Figure 8-1 on page 183 tells us what price Flo must charge to sell that profit-maximizing output. To sell an average of 5.5 packages of her product per week, she must price it at $14.50 (point *k*). The demand curve tells us that this is the only price at which this quantity will be demanded by customers.

Once the profit-maximizing output quantity has been determined with the help of the MR = MC rule, it is easy to find the profit-maximizing price with the help of the demand curve. Just use that curve to find out at what price the *optimal* quantity will be demanded.

GENERALIZATION: THE LOGIC OF MARGINAL ANALYSIS AND MAXIMIZATION

The logic of marginal analysis of profit maximization that we have just studied can be generalized, because essentially the same argument was already used in Chapter 7 and it will recur in a number of the chapters that follow. To avoid having to master the argument each time all over again, it is useful to see how it can be applied in problems other than the determination of the firm's profit-maximizing output.

The general issue is this: Decision makers often are faced with the problem of selecting the magnitude of some variable, such as how much to spend on advertising, or how many bananas to buy, or how many school buildings to construct. Each of these acts brings benefits, so that the larger the number selected by the decision maker, the larger the total benefits that will be derived. But, unfortunately, as larger numbers are selected, the associated costs also grow. The problem is to take the tradeoff properly into account and calculate at what point the net gain—the difference between the total benefit and the total cost—will be greatest. Thus, we have the following general principle:

If a decision is to be made about the quantity of some variable, then to maximize

$$\text{Net benefit} = \text{Total benefit} - \text{Total cost}$$

the decision maker must select a value of the variable at which

$$\text{Marginal benefit} = \text{Marginal cost}$$

For example, if a community were to determine that the marginal benefit from building an additional school was greater than the cost of an additional school, it would clearly be better off if it built another school. On the other hand, if it were building so many schools that the marginal benefit was less than the marginal cost, it would be better off with a more limited construction program. Only if the marginal benefit and cost are as close as possible to being equal will the community have the optimal number of schools.

We will apply this same concept in later chapters. Again and again, we analyze a quantitative decision that brings together both benefits and costs and we always conclude that the optimal decision occurs at the point where the marginal benefit equals the marginal cost. The logic is the same whether we are considering the net gains to a firm, to a consumer, or to society as a whole.

APPLICATION: FIXED COST AND THE PROFIT-MAXIMIZING PRICE

We can now use our analytic framework to offer a surprising insight. Suppose that there is a rise in the firm's fixed cost; say, that Florence Farmer's rent doubles. What will happen to the profit-maximizing price and output? Should she

raise her price to cover the increased cost, or should she produce a larger output even if that requires a fall in price? The surprising answer is: neither!

When a firm's fixed cost increases, its profit-maximizing price and output remain completely unchanged, so long as it pays the firm to stay in business.

In other words, there is nothing that the firm's management can do to offset the effect of the rise in fixed cost. Management must just put up with it. This is surely a case where common sense is not a reliable guide to the right decision.

Why is this so? Remember that, by definition, a fixed cost does not change when output changes. The increase in Flo's fixed costs is the same whether business is slow or booming, whether production is 2 packages of chicken or 200. This is illustrated in Table 8-5, which also reproduces Flo's total profits from Table 8-3. The third column of the table shows that total fixed cost has risen from zero to $5 per week. As a result, total profit is $5 less than it would have been otherwise—no matter what the firm's output. For example, when output is 4 units, we see that total profit falls from $24 (second column) to $19 (last column).

Now, because profit is reduced by the same amount at every output level, whatever output was most profitable before the increase in fixed costs must still be most profitable. In Table 8-5, we see that $22 is the largest entry in the last column, which shows profits after the rise in fixed cost. This highest possible profit is attained, as it was before, when output is between 5 and 6 units. In other words, the firm's profit-maximizing price and quantity will remain exactly as they were before.

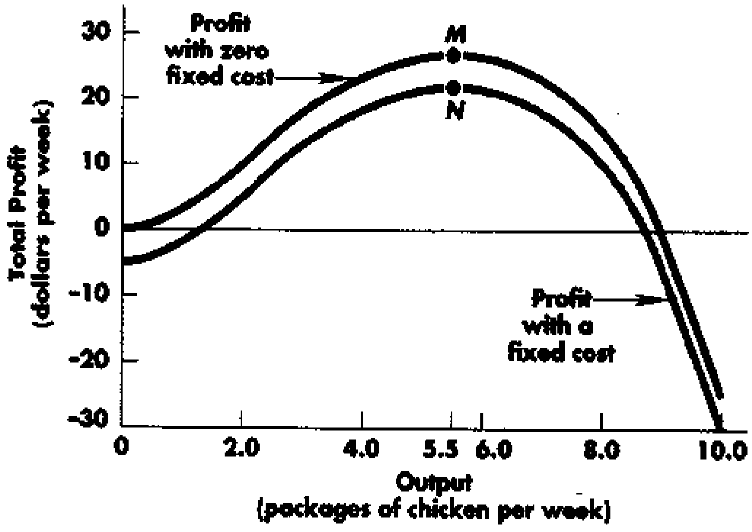
All of this is shown in Figure 8-6, which displays the firm's total profit hill before and after the rise in fixed cost (reproducing Flo's initial profit hill from Figure 8-4). We see that the cost increase simply moves the profit hill straight downward by $5, so the highest point on the hill is just lowered from point *M* to point *N*. But the top of the hill is shifted neither toward the left nor toward the right. It remains at the 110 pounds (or 5.5 packages) output level, just as Table 8-5 indicated.

TABLE 8-5 TOTAL PROFIT BEFORE AND AFTER A RISE IN FIXED COST

Chicken Output (20-lb. packages per week)	Total Profit without Fixed Costs (per week)	Total Fixed Costs (per week)	Total Profit after Fixed Costs (per week)
0	$ 0	$5	−$5.0
1	2.0	5	−3.0
2	10.0	5	5.0
3	18.0	5	13.0
4	**24.0**	**5**	**19.0**
5	27.0	5	22.0
6	27.0	5	22.0
7	23.8	5	18.8
8	16.0	5	11.0
9	0	5	−5.0
10	−25.0	5	−30.0

FIGURE 8-6 FIXED COST DOES NOT AFFECT PROFIT-MAXIMIZING OUTPUT

The graph reproduces Flo's initial profit hill from Figure 8-4 (the black curve labeled "profit with zero fixed cost"). A $5 per week increase in fixed cost shifts the profit hill downward, to the orange curve marked "profit with a fixed cost." But the original point of maximum profit (point *M*) and the new one (point *N*) are at the same output level. This is so because the cost increase pushes the profit hill straight downward.



MARGINAL ANALYSIS IN REAL DECISION PROBLEMS

We can now put the marginal analysis of profit determination to work to unravel the puzzles with which we began this chapter. These are both examples drawn from reality, and reality never works as neatly as a textbook illustration. In particular, neither example involves a mechanical application of the MR = MC rule. However, as these cases show, the underlying reasoning *does* help to deal with real problems.

CASE 1: THE "UNPROFITABLE" CALCULATOR

Our first case study concerned a firm that was apparently losing money on calculator sales because its $12 price was less than the $14.55 average cost which the company's accountants assigned to the product. This $14.55 figure included $10.30 of costs caused directly by manufacture and marketing of the calculators plus a $4.25 per-calculator share of the company's overall general expenses ("overhead"). Accused of deliberately selling below cost in order to drive a competitor out of business, the company turned to marginal analysis to show that this was not true. The calculators were indeed a profitable line of business.

To demonstrate this fact, the company's witness explained that, if the sales were really unprofitable, the company would have been able to raise its net earnings by ceasing production and sale of the calculators. A moment's consideration shows, however, that the opposite would have happened: profits would have decreased if the company gave up its annual sale of 10 million calculators.

The company's revenues would have been reduced by $12 for each of the 10 million units sold—a (marginal) revenue reduction of $120 million. But how much cost would it have saved? The answer is that the cost outlay actually *caused* by the production of each calculator was only the $10.30 in direct cost. The company president would not have been fired if the product were discontinued, and general expenditures on new product research probably would even have

increased. Thus, none of the company's overhead would have been saved by ending calculator production. Rather, the (marginal) saving would have been the direct cost of $10.30 per calculator times the 10 million calculator output—a total saving of $103 million.

Thus, elimination of the product would have reduced total profit by $17 million per year—the $103 million cost saving minus the $120 million in revenue forgone. In other words, continued production of the calculators was not causing losses; on the contrary, it was contributing $17 million in profits every year. The court concluded that this reasoning was correct, and used this conclusion in its decision.

This case illustrates a point that is encountered frequently. The calculator producer was selling its product at a price that appeared not to cover costs but really did. The same sort of issue frequently faces a firm considering the introduction of a new product or the opening of a new branch office. In many such cases, the new operation may not cover *average* costs as measured by standard accounting methods. Yet to follow the apparent implications of those cost figures would amount to throwing away a valuable opportunity to add to the net earnings of the firm (because added revenues exceed added costs) and, perhaps, to contribute to the welfare of the economy. Only *marginal* analysis can reveal whether the contemplated action is really worthwhile.

CASE 2: THE SODA-PRICING PROBLEM

Our second problem dealt with a firm's choice between keeping the price of a brand of soda at $1.50 per six-pack or reducing it to $1.35 when a competitor entered the market. The trouble was that to know what to do, the firm needed to know its demand curve (and hence its marginal revenue curve). However, the firm did not have enough data to determine the shape of its demand curve. How, then, could a rational decision be made?

As we indicated, the debate among the firm's managers finally reached agreement on one point: the price should be cut if, as a result, profits were not likely to decline, that is, if marginal profit were not negative. Fortunately, the data needed to determine whether marginal profit was positive were obtainable. Initial annual sales were 10 million units, and the firm's engineers maintained emphatically that marginal costs were very close to constant at $1.20 per six-pack over the output range in question. Instead of trying to determine the *actual* increase in sales that would result from the price cut, the team of consultants decided to try to determine the *minimum* increase in quantity demanded that would be required to avoid a decrease in profits.

It was clear that the firm needed additional revenue at least as great as the additional cost of supplying the added volume if profits were not to decline; that is, MR had to exceed MC. The consultants knew that sales at the initial price of $1.50 per six-pack were $15 million ($1.50 per unit times 10 million units). Letting Q represent the (unknown) quantity of six-packs that would be sold at the proposed new price of $1.35, the economists compared the added revenue with the added cost of providing the Q new units. Since MC was constant at $1.20 per unit, the added cost amounted to:

$$\text{Added cost} = \$1.20 \times (Q - 10 \text{ million})$$

This was to be compared with the added revenue:

$$\text{Added revenue} = \text{New revenue} - \text{Old revenue} = \$1.35Q - \$15 \text{ million}$$

No loss would result from the price change if the added revenue was greater than or equal to the added cost. The minimum Q necessary to avoid a loss therefore was that at which added revenue equaled added cost, or:

$$1.35Q - 15 \text{ million} = 1.2Q - 12 \text{ million}$$

or

$$0.15Q = 3 \text{ million}$$

This would be true if, and only if, Q, the quantity sold at the lower price, would be:

$$Q = 20 \text{ million units}$$

In other words, this calculation showed that the firm could break even from the 15-cent price reduction only if the quantity of its product demanded rose at least 100 percent (from 10 million to 20 million units). Since past experience indicated that such a rise in quantity demanded was hardly possible, the price reduction proposal was quickly abandoned. Thus, the logic of the MR = MC rule, plus a little ingenuity, enabled the consultants to deal with a problem that at first seemed baffling—even though they had no estimate of marginal revenue.

CONCLUSION: THE FUNDAMENTAL ROLE OF MARGINAL ANALYSIS

We saw in Chapter 7 how marginal analysis helps us to understand the firm's input choices. Similarly, in Chapters 5 and 6, it cast indispensable light on the consumer's purchase decisions. In this chapter, it enabled us to analyze output and pricing decisions. The logic of marginal analysis applies not only to economic decisions by consumers and firms, but also to those of governments, universities, hospitals, and other organizations. In short, the analysis applies to any individual or group that must make economic choices about the use of scarce resources. Thus, one of the most important conclusions that can be drawn from this chapter, a conclusion brought out vividly by the two examples we have just discussed, is:

THE IMPORTANCE OF MARGINAL ANALYSIS

In any decision about whether to expand an activity, it is always the *marginal* cost and *marginal* benefit that are the relevant factors. A calculation based on *average* figures is likely to lead the decision maker to miss all sorts of opportunities, some of them critical.

More generally, if one wants to make *optimal* decisions, *marginal analysis* should be used in the planning calculations. This is true whether the decision applies to a business firm seeking to maximize profit or minimize cost, to a consumer trying to maximize utility, or to a less developed country striving to maximize per-capita output. It applies as much to decisions on input proportions and advertising as to decisions about output levels and prices. Indeed, this is such a general principle of economics that it is one of the **Ideas for Beyond the Final Exam.**

A real-life example far removed from profit maximization will illustrate how marginal criteria are useful in decision making. For some years before women were admitted to Princeton University (and to several other colleges), administrators cited the cost of the proposed change as a major obstacle. They had

decided in advance that any women coming to the university would constitute a net addition to the student body because, for a variety of reasons involving relations with alumni and other groups, it was not feasible to reduce the number of male students. Presumably on the basis of a calculation of average cost, some critics spoke of cost figures as high as $80 million.

To economists it was clear, however, that the relevant figure was the *marginal cost*, the addition to total cost that would result from the admission of the additional students. The women students would, of course, bring to Princeton additional tuition fees (marginal revenues). If these fees were just sufficient to cover the amount that they would *add* to costs, the admission of the women would leave the university's financial picture unaffected.

A careful calculation showed that the admission of women would add far less to the university's financial problems than the *average cost* figures indicated. One reason was that women's course preferences are characteristically different from men's, and hence women frequently select courses that are undersubscribed in exclusively male institutions. Therefore, the admission of 1,000 women to a formerly all-male institution may require fewer additional classes than if 1,000 more men had been admitted.[4] More important, it was found that a number of classroom buildings were underutilized. The cost of operating these buildings was nearly fixed—their total utilization cost would be changed only slightly by the influx of women. The marginal cost for classroom space was therefore almost zero and certainly well below the average cost (the cost per student).

For all of these reasons, it turned out that the relevant marginal cost was much smaller than the figures that had been bandied about earlier. Indeed, this cost was something like a third of the earlier estimates. There is little doubt that this careful marginal calculation played a critical role in the admission of women to Princeton and to some other institutions that made use of the calculations in the Princeton analysis. Subsequent data, incidentally, confirmed that the marginal calculations were amply justified.

THE THEORY AND REALITY: A WORD OF CAUTION

We have now completed two chapters describing how business managers can make optimal decisions. Can you go to Wall Street or Main Street and find executives calculating marginal cost and marginal revenue in order to decide how much to produce? Hardly. Not any more than you can find consumers in stores using marginal analysis in order to decide what to buy. Like consumers, successful business people often rely heavily on intuition and "hunches" that cannot be described by any set of rules. In fact, in a 1993 survey of CEOs conducted by *Inc* magazine, nearly 20 percent of the respondents admitted to using guesswork to price their first product or service.

However, we have not sought a literal *description* of business behavior, but rather a *model* to help us analyze and predict this behavior. The four chapters that we have just completed constitute the core of microeconomics. We will find ourselves repeatedly returning to the principles learned in these chapters.

[4]See Gardner Patterson, "The Education of Women at Princeton," *Princeton Alumni Weekly* 69, September 24, 1968.

SUMMARY

1. A firm can choose the quantity of its product that it wants to sell or the price that it wants to charge. But it cannot choose both because price affects the quantity demanded.
2. In economic theory, we usually assume that firms seek to maximize profits. This should not be taken literally, but rather interpreted as a useful simplification of reality.
3. **Marginal revenue** is the additional revenue earned by increasing sales by 1 unit. Marginal cost is the additional cost incurred by increasing production by 1 unit.
4. Maximum profit requires the firm to choose the level of output at which marginal revenue is equal to marginal cost.
5. Geometrically, the profit-maximizing output level occurs at the highest point on the total profit curve. There the slope of the **total profit** curve is zero, meaning that **marginal profit** is zero.
6. A change in fixed cost will not change the profit-maximizing level of output.
7. It may pay a firm to expand its output if it is selling at a price greater than marginal cost, even if that price happens to be below average cost.
8. **Optimal decisions** must be made on the basis of marginal cost and marginal revenue figures, not average cost and average revenue figures. This is one of the **Ideas for Beyond the Final Exam.**

KEY TERMS

Profit maximization
Satisficing
Optimal decision
Total profit
Economic profit
Total revenue (TR)
Average revenue (AR)
Marginal revenue (MR)
Marginal analysis
Marginal profit

QUESTIONS FOR REVIEW

1. "It may be rational for the management of a firm not to try to maximize profits." Discuss the circumstances under which this statement may be true.
2. Suppose that the firm's demand curve indicates that, at a price of $12 per unit, customers will demand 2 million units of its product. Suppose that management decides to pick *both* price and output; the firm produces 3 million units of its product and prices them at $20 each. What will happen?
3. Suppose that a firm's management would be pleased to increase its share of the market, but if it expands its production the price of its product will fall. Will its profits necessarily fall? Why or why not?
4. Why does it make sense for a firm to seek to maximize *total* profit, rather than to maximize *marginal* profit?
5. A firm's marginal revenue is $112 and its marginal cost is $90. What amount of profit does the firm fail to pick up by refusing to increase output by 1 unit?
6. Calculate average revenue (AR) and average cost (AC) in Table 8-3. How much profit does the firm earn at the output at which AC = AR? Why?
7. A firm's total cost is $800 if it produces 1 unit, $1,400 if it produces 2 units, and $1,800 if it produces 3 units of output. Draw up a table of total, average, and marginal costs for this firm.
8. Draw an average and marginal cost curve for the firm in the preceding question. Describe the relationship between the two curves.
9. A firm has the demand and total cost schedules given in the table below. If it wants to maximize profits, how much output should it produce?

Quantity	Price (dollars)	Total Cost (dollars)
1	6	1.0
2	5	2.5
3	4	6.0
4	3	7.0
5	2	11.0

APPENDIX

THE RELATIONSHIPS AMONG TOTAL, AVERAGE, AND MARGINAL DATA

You may have surmised that there is a close connection between the *average* revenue curve and the *marginal* revenue curve and that there must be a similar relationship between the average cost curve and the marginal cost curve. After all, we derived our total revenue figures from the average revenue and then calculated our marginal revenue figures from the total revenue; a similar relationship applied to costs. In fact:

Marginal, average, and total figures are inextricably bound together. From any one of the three, the other two can be calculated. The relationships among total, average, and marginal figures are exactly the same for *any* variable—such as revenue, cost, or profit—to which the concepts apply.

To illustrate and emphasize the wide applicability of marginal analysis, we switch our example from profits, revenues, and costs to a noneconomic variable. As we are about to see, the same concepts can also be applied to human body weights. We switch to this example because calculation of weights is more familiar to most people than calculation of profits, revenues, or costs, and we can use it to illustrate several fundamental relationships between average and marginal figures.

In Table 8-6, we begin with an empty room. (The total weight of occupants is equal to zero.) A person weighing 100 pounds enters; marginal and average weight are both 100 pounds. If the person is followed by a person weighing 140 pounds (marginal weight equals 140 pounds), the average weight rises to 120 pounds (240/2), and so on.[5]

TABLE 8-6

WEIGHTS OF PERSONS IN A ROOM

Number of Persons in Room	Total Weight (pounds)	Average Weight (pounds)	Marginal Weight (pounds)
0	0	—	
			100
1	100	100	
			140
2	240	120	
			135
3	375	125	
			125
4	500	125	
			100
5	600	120	
			60
6	660	110	

The rule for converting totals to averages, and vice versa, is:

Rule 1a. Average weight equals total weight divided by number of persons.

Rule 1b. Total weight equals average weight times number of persons.

This rule naturally applies equally well to cost, revenue, profit, or any other variable.

We calculate *marginal* weight from *total* weight using the same subtraction process already used to calculate marginal cost and marginal revenue. Specifically:

Rule 2a. The marginal weight of, say, the third person equals the total weight of three people minus the total weight of two people.

For example, when the fourth person enters the room, *total* weight rises from 375 to 500 pounds, and hence the corresponding marginal weight is 500 − 375 = 125 pounds, as is shown in the last column of Table 8-6. We can also do the reverse—calculate total from marginal weight—through an *addition* process.

Rule 2b. The total weight of, say, three people equals the (marginal) weight of the first person who enters the room plus the (marginal) weight of the second person, plus the (marginal) weight of the third person.

You can verify Rule 2b by referring to Table 8-6 which shows that the total weight of three persons, 375 pounds, is indeed equal to 100 + 140 + 135 pounds, the sum of the preceding marginal weights. A similar relation holds for any other total weight figure in the table, as the reader should verify.

[5]Note that in this illustration, "persons in room" is analogous to units of output; "total weight" is analogous to total revenue or cost; and "marginal weight" is analogous to marginal revenue or cost, in the discussions of marginal analysis in the body of the chapter.

In addition to these familiar arithmetic relationships, there are two other useful relationships.

Rule 3. With an exception discussed in the next chapter, the marginal, average, and total figures for the first person must all be equal.

This rule holds because when there is only one person in the room whose weight is X pounds, the average weight will obviously be X, the total weight must be X, and the marginal weight must also be X (since the total must have risen from zero to X pounds). Put another way, when the marginal person is alone, he or she is obviously also the average person, and also represents the totality of all relevant persons.

Our final and very important relationship is:

Rule 4. If marginal weight is lower than average weight, then average weight must fall when the number of persons increases. If marginal weight exceeds average weight, average weight must rise when the number of persons increases. If marginal and average weight are equal, the average weight must remain constant when the number of persons increases.

These three possibilities are all illustrated in Table 8-6. Notice, for example, that when the third person enters the room, the average weight rises from 120 to 125 pounds. That is because this person's (marginal) weight is 135 pounds, which pulls up the average, as Rule 4 requires. Similarly, when the sixth person—who is a 60-pound child—enters the room, the average falls from 120 to 110 pounds because marginal weight, 60 pounds, is below average weight.

It is essential to avoid a common misunderstanding of this rule. It does *not* state, for example, that if the average figure is rising, the marginal figure must be rising. When the average rises, the marginal figure may rise, fall, or remain unchanged. The arrival of two persons, both well above average, will push the average up in two successive steps even if the second new arrival is lighter than the first. We see such a case in Table 8-6, where average weight rises successively from 100 to 120 to 125, while the marginal weight falls from 140 to 135 to 125.

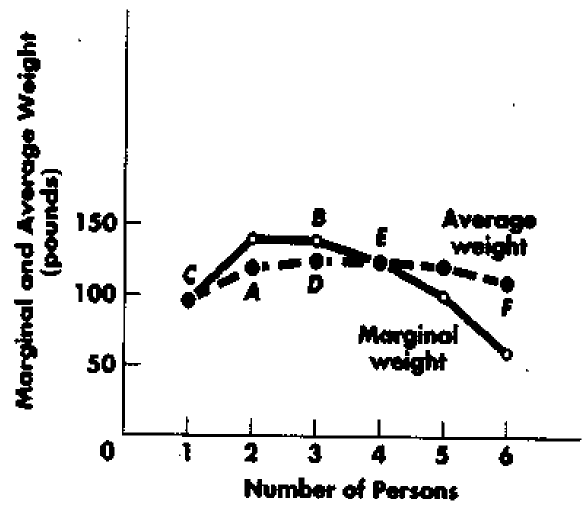
GRAPHICAL REPRESENTATION OF MARGINAL AND AVERAGE CURVES

We have shown how, from a curve of total profit (or total cost or total anything else), one can determine the corresponding marginal figure. We noted several times in the chapter that the marginal value at any particular point is equal to the *slope* of the corresponding total curve at that point. But for some purposes, it is convenient to use a graph that records marginal and average values directly rather than deriving them from the curve of totals.

We can obtain such a graph by plotting the data in a table of average and marginal figures, such as Table 8-6. The result looks like the graph shown in Figure 8-7. Here we have indicated the number of persons in the room on the horizontal axis and the corresponding average and marginal figures on the

FIGURE 8-7 THE RELATIONSHIP BETWEEN MARGINAL AND AVERAGE CURVES

If the marginal curve is above the average curve, the average curve will be pulled upward. Thus, wherever the marginal is above the average, the average must be going upward (orange segment of the curves). The opposite is true where the marginal curve is below the average curve.



vertical axis. The solid dots represent average weights; the small circles represent marginal weights. Thus, for example, point *A* shows that when two persons are in the room, their average weight is 120 pounds, as recorded on the third line of Table 8-6. Similarly, point *B* on the graph represents information provided in the next column of the table, that is, that the marginal weight of the third person who enters the room is 135 pounds. We have connected these points into a marginal curve and an average curve, represented, respectively, by the solid and the broken curves in the diagram. This is the representation of marginal and average values economists most frequently use.

Figure 8-7 illustrates two of our rules. Rule 3 says that, for the first unit, the marginal and average values will be the same; that is precisely why the two curves start out together at point *C*. The graph also depicts Rule 4: between points *C* and *E*, where the average curve is *rising*, the marginal curve lies *above* the average. (Notice, however, that over part of this range, the marginal curve *falls* even though the average curve is rising; Rule 4 says nothing about the rise or fall of the marginal curve.) We see also that over range *EF*, where the average curve is falling, the marginal curve is below the average curve, again in accord with Rule 4. Finally, at point *E*, where the average curve is neither rising nor falling, the marginal curve meets the average curve: the average and marginal weights are equal at that point.

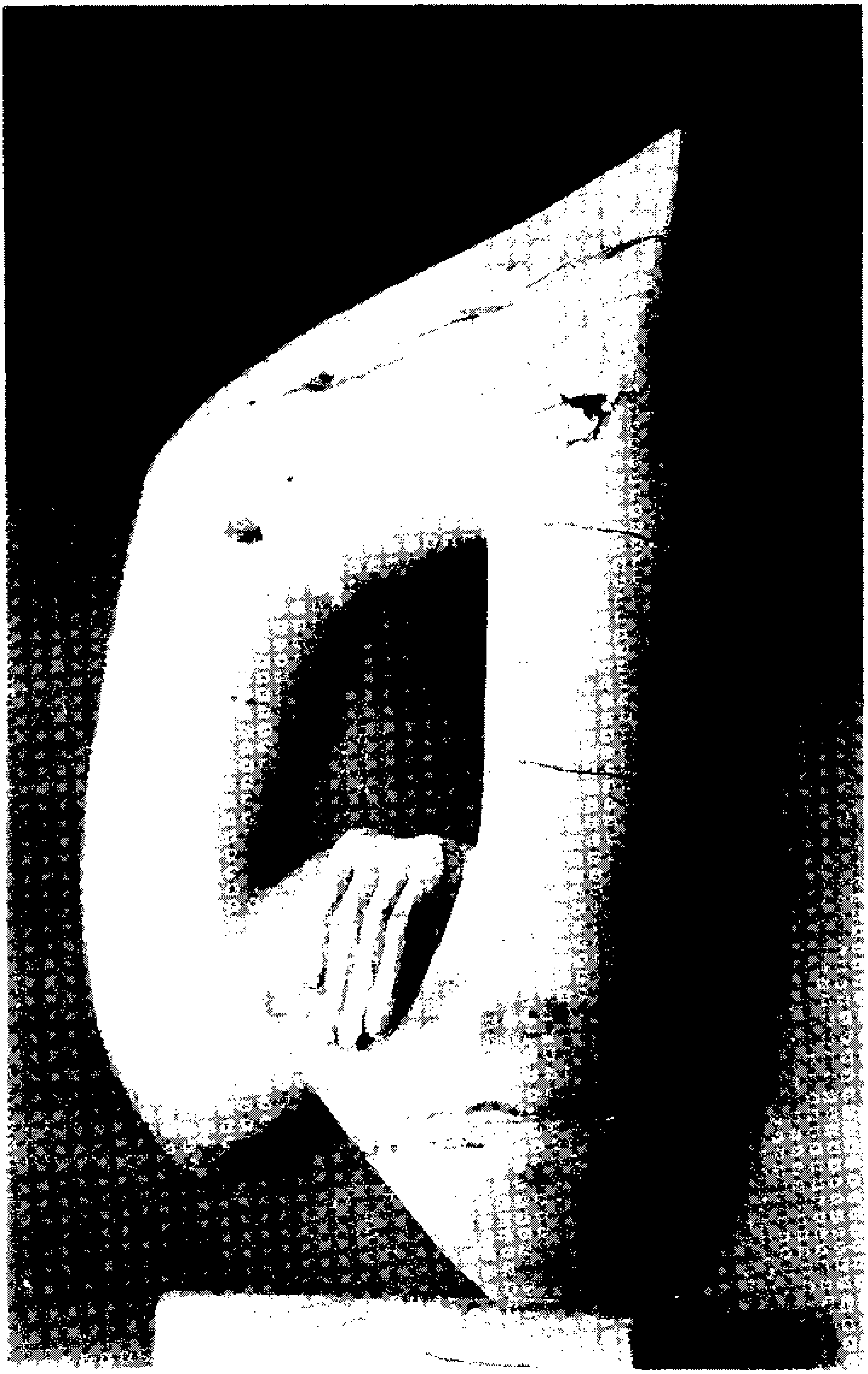
QUESTIONS FOR REVIEW

1. Suppose that the following table is your record of exam grades in your Principles of Economics course:

Exam Date	Grade	Comment
September 30	65	A slow start
October 28	75	A big improvement
November 26	90	Happy Thanksgiving!
December 13	85	Slipped a little
January 24	95	A fast finish!

   Use these data to make up a table of total, average, and marginal grades for the five exams.

2. From the data in your table, illustrate each of the rules mentioned in this appendix. Be sure to point out an instance where the marginal grade falls but the average grade rises.



PART III

THE MARKET SYSTEM: VIRTUES AND VICES

CHAPTER 9

THE FIRM AND THE INDUSTRY UNDER PERFECT COMPETITION

Competition . . . brings about the only . . . arrangement of social production which is possible. . . . [Otherwise] what guarantee [do] we have that the necessary quantity and not more of each product will be produced, that we shall not go hungry in regard to corn and meat while we are choked in beet sugar and drowned in potato spirit, that we shall not lack trousers to cover our nakedness while buttons flood us in millions.
Friedrich Engels

Besides the consumer demands and the business costs that we studied in the last four chapters, the decisions of firms also depend on the number, sizes, and behavior of other firms in the industry. The strength of the competition faced by a company can profoundly affect its pricing, its output decisions, and its input purchases. Strong competitive pressures, sometimes taking subtle forms, can severely limit the freedom of choice of management in setting prices. In the process, those pressures can protect the interests of consumers. Even giant corporations with few domestic rivals may find themselves under this sort of pressure. In recent years, for example, many American companies have faced stiffening competition from foreign companies, including Japanese, German, and Swedish firms. This chapter and the four chapters that follow it analyze some of the forms that competition—or its absence—can take, along with some of the implications of competition for the general welfare of consumers and producers.

Industries differ dramatically in the number and typical sizes of their firms. Some industries, such as fishing, contain a great many very small firms; others, like automobile manufacturing, are composed of a few industrial giants. This chapter deals with a very special type of market structure—called *perfect competition*— in which firms are numerous and small. We begin by comparing alternative market forms and defining perfect competition precisely. We then use the tools acquired in Chapter 8 to analyze the behavior of the perfectly competitive firm. Next, we consider *all* the firms in an industry as a group, and we investigate how developments in the industry affect the individual firms.

TWO PRACTICAL PUZZLES

We begin this chapter, once again, with two puzzling but important applications:

ISSUE 1: POLLUTION-REDUCTION INCENTIVES THAT INCREASE POLLUTION

Many economists and others who are concerned with the environment believe that cleaner air and water can be achieved cheaply and effectively by requiring polluters to pay for the damages they cause. Firms will have an incentive to reduce pollution if this helps them to escape fines. (See Chapter 21 for more details.)

Yet people often view pollution charges as taxes, and that can make the charges political poison. Some politicians, reasoning that they can make a donkey move as effectively with a carrot as with a stick, have proposed *paying* firms to cut down their emissions of pollutants.

There is at least theoretical evidence that this system of bribes (or subsidies) does work, *at least up to a point*. Individual polluting firms will, indeed, respond by cutting pollution. But, over time, it turns out that society will end up with more pollution than before! Payments to the firms actually make the pollution problem worse. How is it possible that each firm pollutes less, and yet total pollution increases?

ISSUE 2: CAN FIRMS SHIFT THE BURDEN OF TAXATION TO CONSUMERS?

It is widely believed that a tax on business is not borne by the firm, but rather is passed on to consumers in the form of higher prices. "Firms don't pay taxes, people do." Yet whenever lawmakers propose a tax on some product, lobbyists for the targeted industry invariably show up at the legislature, arguing against the program. If firms can simply pass on the tax to the consuming public, why should they waste money and effort fighting it?

By the end of the chapter, the analysis will have supplemented your common sense sufficiently to supply the required insights into these two problems.

VARIETIES OF MARKET STRUCTURE: A SNEAK PREVIEW

A **MARKET** refers to the set of all sale and purchase transactions that affect the price of some commodity.

So far, we have talked only about firms in general without worrying about the sorts of markets in which they operate. It will be helpful, next, to explain clearly what is meant by the word **market**. Economists do not reserve the term to denote only an organized exchange operating in a well-defined physical location. In its more general and abstract usage, the word *market* refers to a set of sellers and buyers whose activities affect the price at which a *particular commodity* is sold. For example, two separate sales of General Motors stock in different parts of the country may be considered as taking place on the same market, while sales of bread and carrots in adjacent stalls of a market square may, in our sense, occur on totally different markets.

In this chapter and the next few, we will see that the type of market in which the firm operates makes a great deal of difference in its behavior. Under some market forms, for example, the firm has no control over its price. In others, the firm has the power to adjust its price in a way that adds to its profit and which, in the opinion of some, constitutes exploitation of consumers.

Economists distinguish among different kinds of markets according to (1) how many firms they include, (2) whether the products of the different firms are identical or somewhat different, and (3) how easy it is for new firms to enter the markets. *Perfect competition* is at one extreme (many small firms selling an identical

product), while *pure monopoly* (a single firm) is at the other. In between are hybrid forms—called *monopolistic competition* (many small firms each selling slightly different products) and *oligopoly* (a few large rival firms)—that share some of the characteristics of both perfect competition and monopoly.

Perfect competition is far from the typical market form in the U.S. economy. Indeed, it is quite rare. Pure monopoly—literally *one* firm—is also infrequently encountered. Most of the products you buy are no doubt supplied by oligopolies or monopolistic competitors—terms that we will define precisely in Chapter 12.

PERFECT COMPETITION DEFINED

PERFECT COMPETITION occurs in an industry when that industry is made up of many small firms producing homogeneous products, information is perfect, and there is no impediment to entry or exit of firms.

You can appreciate just how special perfect competition is once we provide a comprehensive definition. A market is said to operate under **perfect competition** when the following four conditions are satisfied:

1. *Numerous small firms and customers.* The market has so many buyers and sellers that each one constitutes a negligible portion of the whole, so small, in fact, that its decisions have no effect on price. This requirement rules out trade associations or other collusive arrangements strong enough to affect price.
2. *Homogeneity of product.* The product offered by any seller is identical to that supplied by any other seller. (For example, wheat of a given grade is a homogeneous product; different brands of toothpaste are not.) Because the product is homogeneous, consumers do not care from which firm they buy.
3. *Freedom of entry and exit.* New firms desiring to enter the market face no impediments that the existing firms can avoid. Similarly, if production and sale of the good proves unprofitable, there are no barriers preventing firms from leaving the market.
4. *Perfect information.* Each firm and each customer is well informed about the available products and their prices. They know whether one supplier is selling at a lower price than another.

These are obviously exacting requirements that are met infrequently in practice. One example might be a market for common stocks. There are literally millions of buyers and sellers of AT&T stock, all of the shares are exactly alike, anyone who wishes can enter the market easily, and most of the relevant information is readily available in the daily newspaper. Many farming and fishing industries also approximate perfect competition. But it is hard to find many other examples. Our interest in perfect competition is surely not for its descriptive realism.

Why, then, do we spend time studying perfect competition? The answer takes us back to the central theme of this book. It is under perfect competition that the market mechanism performs best. If we want to learn what markets do well, we can put the market's best foot forward by beginning with perfect competition.

As Adam Smith suggested some two centuries ago, perfectly competitive firms use society's scarce resources with maximum efficiency. Also, as Friedrich Engels (the closest friend and coauthor of Karl Marx) suggested in the opening quotation of this chapter, only perfect competition can ensure that the economy turns out just those varieties and relative quantities of goods that match the preferences of consumers. By studying perfect competition, we can learn just what an *ideally functioning* market system can accomplish. This is the topic of the

present chapter and the next one. Then, in Chapters 11 and 12, we will consider other market forms and see how they deviate from the perfectly competitive ideal. Still later chapters (especially Chapter 13 and the chapters in Parts IV and V) will examine many important tasks that the market does not perform at all well, even under perfect competition. These chapters combined should provide a balanced assessment of the virtues and vices of the market mechanism.

THE COMPETITIVE FIRM AND ITS DEMAND CURVE

To discover what happens in a perfectly competitive market, we must deal separately with the behavior of *individual firms* and the behavior of the *industry* that is constituted by those firms. One basic difference between the firm and the industry under competition relates to *pricing*:

Under perfect competition, the firm has no choice but to accept the price that has been determined in the market. We say that it is a *price taker*.

That each firm in a perfectly competitive market has no control over its price follows from the definition of perfect competition. The presence of a vast number of competitors, each offering identical products, forces each firm to meet but not exceed the price charged by the others. But while the individual firm has no influence over price under perfect competition, the industry does. This influence is not conscious or planned. It happens spontaneously through the impersonal forces of supply and demand, as we observed in Chapter 4.

With two important exceptions, the analysis of the behavior of the firm under perfect competition is exactly as described in Chapters 7 and 8. The two exceptions are the special shape of the competitive firm's demand curve and the effects of freedom of entry and exit on the firm's profits. We will consider them in turn, beginning with the demand curve.

In Chapter 8, we always assumed that the firm's demand curve sloped downward; if a firm wished to sell more (without increasing its advertising or changing its product specifications), it had to reduce the price of its product. The competitive firm is an exception to this general principle.

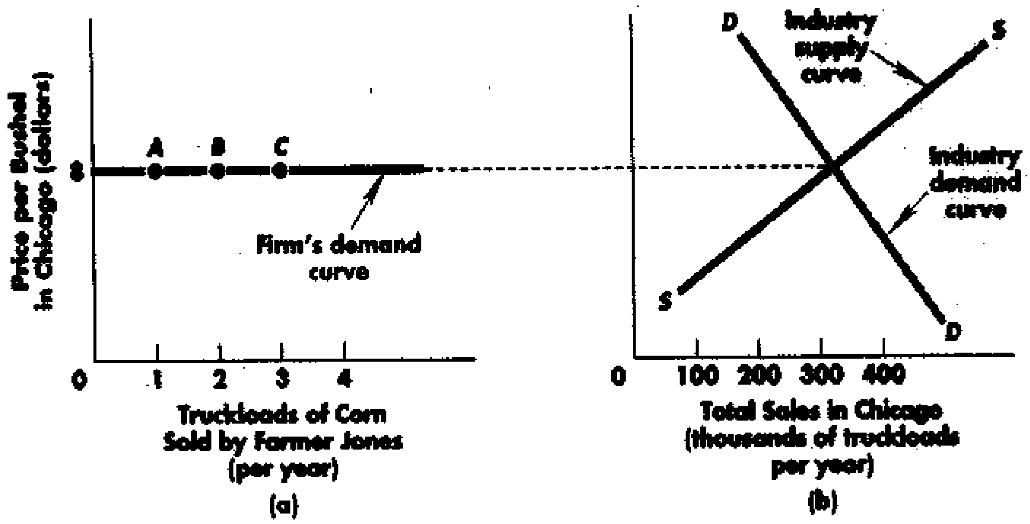
A perfectly competitive firm has a horizontal demand curve. This means that it can sell as much as it wants at the prevailing market price. It can double or triple its sales without any reduction in the price of its product.

How is this possible? The answer is that the competitive firm is so insignificant relative to the market as a whole that it has absolutely no influence over price. The farmer who sells his corn through an exchange in Chicago must accept the current quotation his broker reports to him. Because there are thousands of farmers, the Chicago price per bushel will not budge because Farmer Jones decides he doesn't like the price and holds back a truckload for storage.

Thus, the demand curve for Farmer Jones's corn is as shown in Figure 9-1(a); the price he is paid in Chicago will be $8 per bushel whether he sells one truckload (point *A*) or two (point *B*) or three (point *C*). That is because price is determined by the *industry's* supply and demand curves shown in the right-hand portion of the graph (Panel b).

Notice that, in the case of perfect competition, the downward-sloping industry demand curve in Figure 9-1(b) leads to the horizontal demand curve for the individual firm in Figure 9-1(a). Thus, the firm's demand need not resemble the demand curve for the industry.

FIGURE 9-1 DEMAND CURVE FOR A FIRM UNDER PERFECT COMPETITION



Under perfect competition, the size of the output of a firm is so small a portion of the total industry output that it cannot affect the market price of the product. Even if the firm's output increases many times, market price remains $8, where it is set by industry supply and demand.

SHORT-RUN EQUILIBRIUM OF THE PERFECTLY COMPETITIVE FIRM

We already have sufficient background to study the decisions of a firm operating in a perfectly competitive market. Recall from Chapter 8 that profit maximization requires the firm to pick an output level that makes its *marginal cost equal to its marginal revenue*: MC = MR. The only feature that distinguishes the profit-maximizing equilibrium of the competitive firm from that of any other type of firm is its horizontal demand curve.

Because the demand curve (the average revenue curve) is horizontal, the competitive firm's marginal revenue curve is a horizontal straight line that coincides with its demand curve; hence, MR = Price (*P*). It is easy to see why this is so. If the price does not depend on how much the firm sells (which is what a horizontal demand curve means), then each *additional* unit sold brings in an amount of revenue (the *marginal* revenue) exactly equal to the market price. So marginal revenue always equals price under perfect competition because the firm is a price taker.

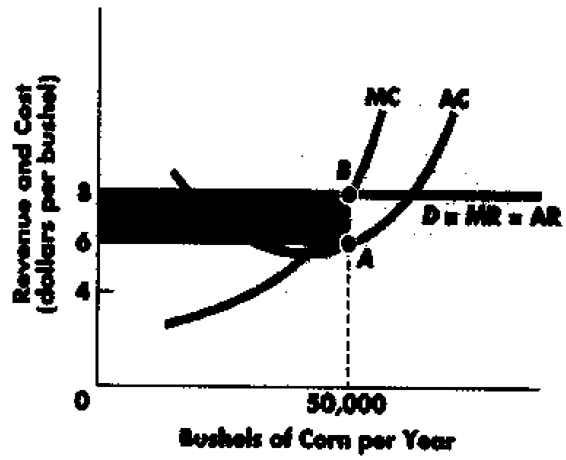
Once we know the shape and position of a firm's marginal revenue curve, we can use this information and the marginal cost curve to determine its optimal output and profit, as shown in Figure 9-2. As usual, the profit-maximizing output is that at which MC = MR (point *B*). This competitive firm produces 50,000 bushels per year—the output level at which MC and MR are both are equal to the market price, $8. Thus:

Because it is a price taker, the *equilibrium* of a profit-maximizing firm in a perfectly competitive market must occur at an output level at which marginal cost is equal to price. This is because a horizontal demand curve makes price and MR equal and, therefore, both are equal to marginal cost. In symbols:

$$MC = P$$

FIGURE 9-2 SHORT-RUN EQUILIBRIUM OF THE COMPETITIVE FIRM

The profit-maximizing firm will select the output (50,000 bushels per year) at which marginal cost equals marginal revenue (point *B*). The demand curve, *D*, is horizontal because the firm's output is too small to affect market price; thus, it is also the marginal revenue curve. In the short run, demand may be either high or low in relation to cost. Therefore, each unit the firm sells may return a profit (*AB*) or a loss.



This idea is illustrated in Table 9-1, which gives the firm's total and marginal revenue, total and marginal cost, and total profit for different output quantities. We see from the last column that total profit is maximized at an output of about 50,000 bushels where total profit is $100,000. An increase in output from 40,000 to 50,000 bushels incurs a marginal cost ($70,000) which is approximately equal to the corresponding marginal revenue ($80,000), confirming that 50,000 bushels is the profit-maximizing output.[1]

TABLE 9-1 REVENUES, COSTS, AND PROFITS OF A COMPETITIVE FIRM (THOUSANDS OF DOLLARS)

Total Quantity (thousands of bushels)	Total Revenue	Marginal Revenue	Total Cost	Marginal Cost	Total Profit
0	0				
		80			
10	80		85		−5
		80		65	
20	160		150		10
		80		30	
30	240		180		60
		80		50	
40	320		230		90
		80		70	
50	400		300		100
		80		150	
60	480		450		30
		80		250	
70	560		700		−140

[1] To calculate marginal costs and marginal revenues accurately, we should increase output one bushel at a time instead of proceeding in leaps of 10,000 bushels. But that would require too much space! In any event, our failure to make a more careful calculation in terms of individual bushels explains why we are unable to find the output at which MR and MC are exactly equal.

SHORT-RUN PROFIT: GRAPHIC REPRESENTATION

Our analysis so far tells us how the firm can pick the output that maximizes its profit. But even if it succeeds in doing so, the firm may conceivably find itself in trouble. If the demand for its product is weak or its costs are high, even the firm's most profitable option may lead to a loss. To determine whether the firm is making a profit or incurring a loss, we must compare *total* revenue (TR = $P \times Q$) with *total* cost (TC = AC $\times$ Q). Since Q is common to both of these, that is equivalent to comparing P with AC.

The firm's profit can therefore be shown in Figure 9-2, which includes the firm's *average cost* (AC) curve. By definition, profit per unit of output is revenue per unit (P) minus cost per unit (AC). We see in Figure 9-2 that average cost at 50,000 bushels per year is only $6 per bushel (point *A*), while *average revenue* (AR) is $8 per bushel (point *B*). The firm is making a profit of AR − AC = $2 per bushel, which appears in the graph as the vertical distance between points *A* and *B*.

Notice that in addition to showing the *profit per unit*, the graph can be used to show the firm's *total profit*. Total profit is the profit per unit ($2 in this example) times the number of units (50,000 per year). Therefore, total profit is represented by the *area* of the shaded rectangle whose height is the profit per unit ($2) and whose width is the number of units (50,000).[2] In this case, profits are $100,000 per year. In general, total profit at any output is the area of the rectangle whose base equals the level of output and whose height equals AR − AC.

The MC = P condition gives us the output that maximizes the perfectly competitive firm's profit. It does not, however, tell us whether the firm is making a profit or incurring a loss. To determine this, we must compare price with average cost.

THE CASE OF SHORT-TERM LOSSES

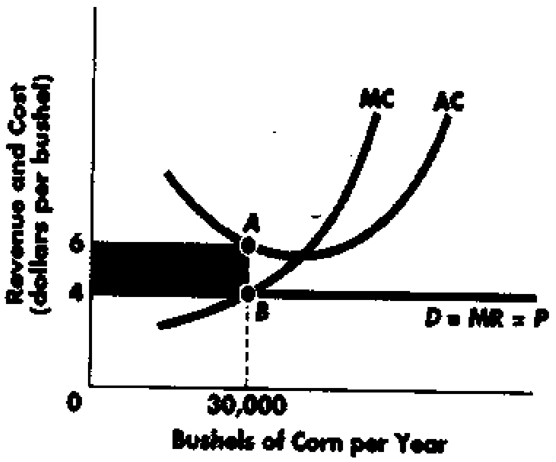
The market is obviously treating the farmer in the graph rather nicely. But what if the market were not so generous in its rewards? What if, for example, the market price were only $4 per bushel instead of $8? Figure 9-3 shows the equilibrium of the firm under these circumstances. The firm still maximizes profits by producing the level of output at which marginal cost is equal to price—point *B* in the diagram. But this time "maximizing profits" really means minimizing losses.

At the optimal level of output (30,000 bushels per year), average cost is $6 per bushel (point *A*), which exceeds the $4 per bushel price (point *B*). The firm is therefore running a loss of $2 per bushel times 30,000 bushels, or $60,000 per year. This loss, which is represented by the area of the shaded rectangle in Figure 9-3, is the best the firm can do. If it selected any other output level, its loss would be even greater.

[2]Recall that the formula for the area of a rectangle is area = height × width.

FIGURE 9-3 SHORT-RUN EQUILIBRIUM OF THE COMPETITIVE FIRM WITH A LOWER PRICE

In this diagram, the cost curves are the same as in Figure 9-2 but the demand curve (*D*) has shifted down to a market price of $4 per bushel. The firm still does the best it can by setting MC = *P* (point *B*). But since its average cost at 30,000 bushels per year is $6 per bushel, it runs a loss (shown by the shaded rectangle).



SHUTDOWN AND BREAK-EVEN ANALYSIS

There is, however, a limit to the loss the firm can be forced to accept. If losses get too big, the firm can simply go out of business. To understand the logic of the choice between shutting down and remaining in operation, we must return to the distinction between costs that are sunk and those that are variable in the short run. Recall from Chapter 7 that costs are sunk if the firm cannot escape them in the short run, either because of a contract (say, with the landlord or the union) or because it has already bought the item whose cost is sunk (for example, a machine).

If the firm stops producing, its revenue will fall to zero. Its short-run variable costs will also fall to zero. But its sunk costs will remain. If the firm is losing money, sometimes it will be better off continuing to operate until its obligations to pay sunk costs expire; sometimes it will do better by shutting down and producing nothing. Two rules govern the decision:

Rule 1. The firm will make a profit if total revenue (TR) exceeds total cost (TC). In that case, it should not plan to shut down either in the short run or in the long run.

Rule 2. The firm should continue to operate in the short run if TR exceeds total short-run variable cost (TVC). It should nevertheless plan to close in the long run if TR is less than TC.

The first rule is obvious. If the firm's revenues cover its total costs, then it does not lose money. The second rule is a bit more subtle. Suppose that TR is less than TC. If our unfortunate firm continues in operation, it will lose the difference between total cost and total revenue, that is:

$$\text{Loss if the firm stays in business} = TC - TR$$

However, if the firm stops producing, both its revenues and short-run variable costs become zero, leaving only the *sunk* costs to be paid:

$$\text{Loss if the firm shuts down} = \text{Sunk costs} = TC - TVC$$

Hence, it is best to keep operating as long as:

$$TC - TR < TC - TVC$$

or

$$TVC < TR$$

That is Rule 2.

We can illustrate Rule 2 by the two cases in Table 9-2. Case A deals with a firm that loses money but is better off staying in business in the short run. If it shuts down, it will lose its entire $60,000 sunk cost. But if it continues to operate, total revenue of $100,000 exceeds total variable cost (TVC = $80,000) by $20,000. That means continuing operation contributes $20,000 toward meeting sunk costs and reduces losses to $40,000. In Case B, on the other hand, it pays the firm to shut down because continued operation only adds to its losses. If the firm operates, it will lose $90,000 (the last entry in Table 9-2), whereas if it shuts down, it will lose only the $60,000 in sunk costs which it must pay whether it operates or not.

We can also analyze the shutdown decision graphically. In Figure 9-4, the firm will run a loss whether the price is P_1, P_2, or P_3, because none of these prices is high enough to reach the minimum level of average cost (AC). We can show the *lowest* price that keeps the firm from shutting down by introducing one more short-run cost curve: the average variable cost (AVC) curve. Why is this curve relevant? Because, as we have just seen, it pays the firm to remain in operation if its total revenue (TR) exceeds its total short-run variable cost (TVC). If we divide both TR and TVC by quantity (Q), we get $TR/Q = P$ and $TVC/Q = AVC$, so we can state this condition equivalently as the requirement that price must exceed AVC. The conclusion is:

The firm will produce nothing unless price lies above the minimum point on the AVC curve.

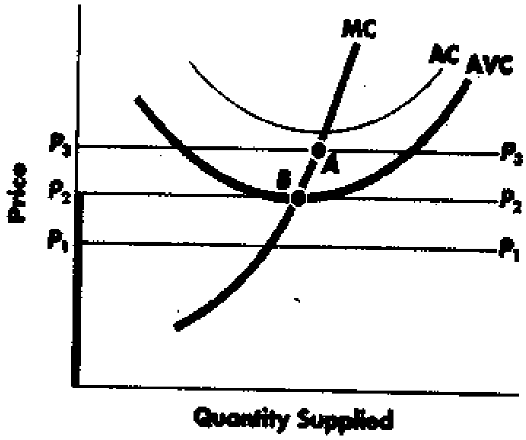
In Figure 9-4, price P_1 is below the minimum average variable cost. With this price, the firm cannot even cover its variable costs and is better off shutting down (producing zero output). Price P_3 is higher. While the firm still runs a loss if it sets $MC = P$ at point A (because AC exceeds P_3), it is at least covering its short-run variable costs, and so it pays to keep operating in the short run. Price P_2 is the borderline case. If the price is P_2, the firm is indifferent between shutting down and staying in business and producing at a level where $MC = P$ (point

TABLE 9-2 THE SHUTDOWN DECISION

	Case A	Case B
	(thousands of dollars)	
Total revenue (TR)	$100	$100
Total variable cost (TVC)	80	130
Sunk cost	60	60
Total cost (TC)	140	190
Loss if firm shuts down (= sunk cost)	60	60
Loss if firm does not shut down	40	90

FIGURE 9-4 SHUTDOWN ANALYSIS

At a price as low as P_1, the firm cannot even cover its short-run average variable costs, and it is better off shutting down entirely. At a price as high as P_3, the firm selects point *A* but operates at a loss (because P_3 is below AC). However, it is more than covering its average variable costs (since P_3 exceeds AVC), so it pays to keep producing. Price P_2 is the borderline case. With this price, the firm selects point *B* and is indifferent between shutting down and staying open.



B). P_2 is thus the *lowest* price at which the firm will produce anything. As we see from the graph, P_2 corresponds to the minimum point on the AVC curve.

THE SHORT-RUN SUPPLY CURVE OF THE COMPETITIVE FIRM

Without realizing it, we have now derived the supply curve of the competitive firm in the short run. Why? Recall that a supply curve summarizes in a graph the answers to questions such as, "If the price is so and so, how much output will the firm offer for sale?" We have now discovered that there are two possibilities, as indicated by the thick orange line in Figure 9-4.

1. In the short run, if the price exceeds the minimum AVC, it pays a competitive firm to produce the level of output that equates MC and *P*. Thus, for any price above point *B*, we can read the corresponding quantity supplied from the firm's MC curve.
2. If the price falls below the minimum AVC, then it pays the firm to produce nothing. Quantity supplied falls to zero.

Putting these two observations together, we conclude that:

The short-run supply curve of the perfectly competitive firm is the portion of its marginal cost curve that is above the point where it intersects the average (short-run) variable cost curve, that is, above the minimum level of AVC. If price falls below this level, the firm's quantity supplied drops to zero.

THE SHORT-RUN SUPPLY CURVE OF THE COMPETITIVE INDUSTRY

Having completed the analysis of the competitive *firm's* supply decision, we turn our attention next to the competitive *industry*. Again we need to distinguish between the short run and the long run, but the distinction is different here. The short run for the *industry* is defined as a period of time too brief for new firms to enter the industry or for old firms to leave, so the number of firms is fixed. By contrast, the long run for the industry is a period of time long enough for

any firm that so desires to enter or leave. In addition, in the long run each firm in the industry can adjust its output to its own long-run costs.[3] We begin our analysis of industry equilibrium in the short run.

With the number of firms fixed, it is a simple matter to derive the supply curve of the competitive industry from those of the individual firms. At any given price, we simply *add up* the quantities supplied by each of the firms to arrive at the industrywide quantity supplied. For example, if each of 1,000 identical firms in the corn industry supplies 45,000 bushels when the price is $6 per bushel, then the quantity supplied by the industry at a $6 price will be 45,000 bushels per firm × 1,000 firms = 45 million bushels.

This process of deriving the *market* supply curve from the *individual* supply curves of firms is perfectly analogous to the way we derived the *market* demand curve from the *individual* demand curves of consumers in Chapter 6. Graphically, what we are doing is *summing the individual supply curves horizontally*, as illustrated in Figure 9-5. At a price of $6, each of the 1,000 identical firms in the industry supplies 45,000 bushels [point *c* in Panel (a)], so the industry supplies 45 million bushels [point C in Panel (b)]. At a price of $8, each firm supplies 50,000 bushels [point *e* in Panel (a)], and so the industry supplies 50 million bushels [point E in Panel (b)]. We can carry out similar calculations for any other price.

The supply curve of the competitive industry in the short run is derived by *summing* the short-run supply curves of all the firms in the industry *horizontally*.

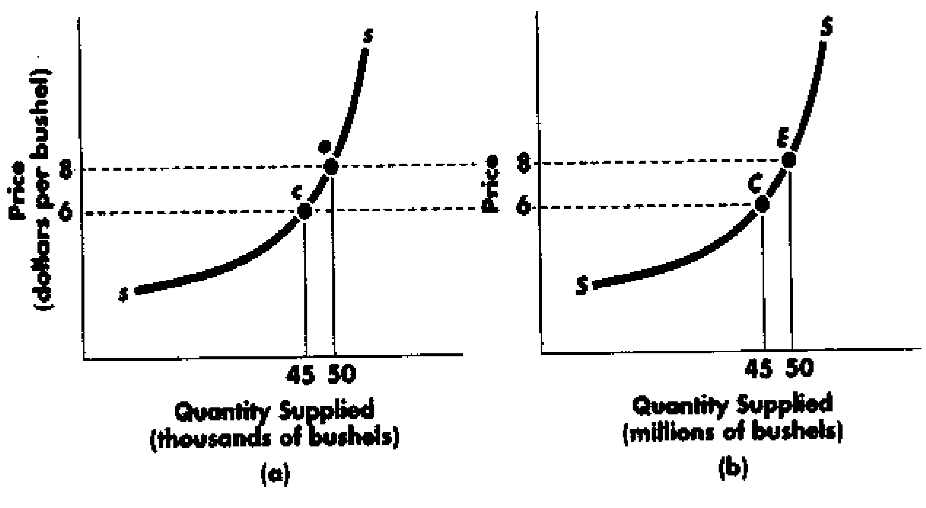
This adding-up process indicates, incidentally, that the supply curve of the industry will shift to the right whenever a new firm enters the industry.

Notice that if the short-run supply curves of individual firms are upward sloping, then the short-run supply curve of the competitive industry will be upward sloping, too. We have seen that the firm's supply curve is its marginal cost curve (above the level of minimum average variable cost), so it follows that rising marginal costs lead to an upward-sloping short-run *industry* supply curve.

FIGURE 9-5

DERIVATION OF THE INDUSTRY SUPPLY CURVE FROM THE SUPPLY CURVES OF THE INDIVIDUAL FIRMS

In this hypothetical industry of 1,000 identical firms, each individual firm has the supply curve *ss* in Panel (a). For example, quantity supplied is 45,000 bushels when the price is $6 per unit (point *c*). By *adding up* the quantities supplied by each firm at each possible price, we arrive at the industry supply curve, *SS* in Panel (b). For example, at a unit price of $6, total quantity supplied by the industry is 45 million units (point *C*).



[3]The relationship between short-run and long-run cost curves for the firm was discussed in Chapter 7, pages 160–161.

INDUSTRY EQUILIBRIUM IN THE SHORT RUN

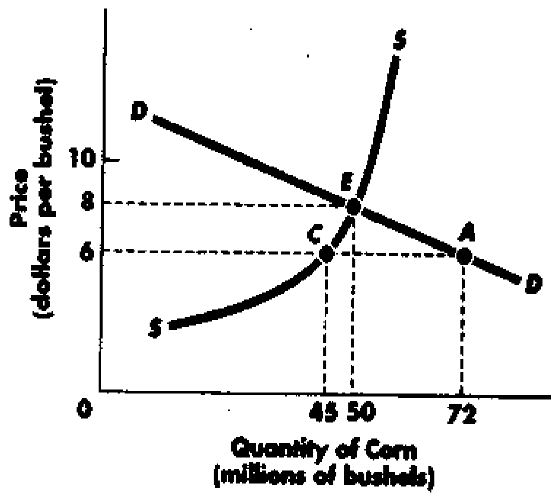
Now that we have derived the industry supply curve, we need only add a market demand curve to determine the price and quantity that will emerge. We do this for our illustrative corn industry in Figure 9-6, where the orange industry supply curve, carried over from Figure 9-5(b), is *SS* and the demand curve is *DD*. Note that for the competitive industry, unlike the competitive firm, the demand curve is normally downward sloping. Why? Each firm by itself is so small that, if it alone were to double its output, the effect would hardly be noticeable. But if *every* firm in the industry were to expand its output, that would make a substantial difference. Customers can be induced to buy the additional quantities arriving at the market only if the price of the good falls.

Point *E* is the equilibrium point for the competitive industry, because only at a price of $8 are sellers willing to offer exactly the amount that consumers want to purchase (in this case, 50 million bushels).

Should we expect price actually to reach, or at least to approximate, this equilibrium level? The answer is yes. To see why, we must consider what happens when price is not at its equilibrium level. Suppose that the price is lower, say, $6. This low price will stimulate customers to buy more; it will also lead firms to produce less than they would at a price of $8. Our diagram confirms that at a price of $6, quantity supplied (45 million bushels) is lower than quantity demanded (72 million bushels). Thus, unsatisfied buyers will probably offer to pay higher prices, which will force the price *upward* in the direction of its equilibrium value, $8.

Similarly, if we begin with a price higher than the equilibrium price, we may readily verify that quantity supplied will exceed quantity demanded. Under these circumstances, frustrated sellers are likely to reduce their prices, so price will be forced downward. In the circumstances depicted in Figure 9-6, then, there

FIGURE 9-6 SUPPLY-DEMAND EQUILIBRIUM OF A COMPETITIVE INDUSTRY



The only equilibrium combination of price and quantity is a price of $8 and a quantity of 50 million bushels, at which the supply curve, *SS*, and the demand curve, *DD*, intersect (point *E*). At a lower price such as $6, quantity demanded (72 million bushels, as shown by point *A* on the demand curve) will be higher than the 45-million-bushel quantity supplied (point *C*). Thus, the price will be driven back up toward the $8 equilibrium. The opposite will happen at a price such as $10, which is above equilibrium.

is in effect a magnet at the equilibrium price of $8 that will pull the actual price in its direction, if for some reason the actual price starts out at some other level.

In practice, there are few cases in which competitive markets, over a long period of time, seem not to have moved toward equilibrium prices. Matters eventually appear to work out as depicted in Figure 9-6. Of course, numerous transitory influences can jolt any real-world market away from its equilibrium point—a workers' strike that cuts production, a sudden change in consumer tastes, and so on.

Yet, as we have just seen, there are powerful forces that push prices back toward equilibrium—toward the level at which the supply and demand curves intersect. These forces are of fundamental importance for economic analysis. If there were no such forces, prices in the real world would bear little resemblance to equilibrium prices, and there would be little reason to study supply-demand analysis. Fortunately, the required equilibrating forces do exist.

INDUSTRY AND FIRM EQUILIBRIUM IN THE LONG RUN

The equilibrium of a competitive industry in the long run may differ from the short-run equilibrium that we have just studied for two reasons. First, the number of firms in the industry (1,000 in our example) is not fixed in the long run. Second, as we saw in Chapter 7 (pages 160–161), the firm can vary its plant size and make other changes in the long run that were prevented by temporary commitments in the short run. Hence, the firm's (and the industry's) long-run cost curves are not the same as the short-run cost curves.

What will lure new firms into the industry or repel old ones? In a word, profits. Remember that when a firm selects its optimal level of output by setting MC = *P* it may wind up with either a profit or a loss. Such profits or losses must be *temporary* for a competitive firm, because the freedom of new firms to enter the industry or of old firms to leave it will, in the long run, eliminate them.

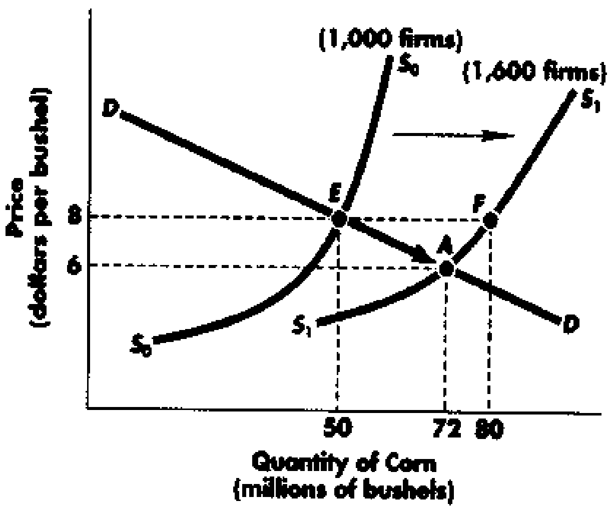
Suppose that firms in the industry earn very high profits. Then new companies will find it attractive to enter the business, and expanded production will force the market price to fall from its initial level. Why? Recall that the industry supply curve is the horizontal sum of the supply curves of individual firms. Under perfect competition, new firms can enter the industry *on the same terms as existing firms.* This means that new entrants will have the *same* individual supply curves as old firms. If the market price did not fall, entry of new firms would lead to an increased number of firms with no change in output *per firm.* Consequently, the total quantity supplied on the market would be higher, and it would exceed quantity demanded. But, of course, this means that entry of new firms *must* push the price down.

Figure 9-7 shows how the entry process works. In this diagram, the demand curve *DD* and the original (short-run) supply curve S_0S_0 are carried over from Figure 9-6. The entry of new firms seeking high profits *shifts the industry's short-run supply curve outward to the right,* to S_1S_1. The new market equilibrium is at point *A* (rather than at point *E*), where price is $6 per bushel and 72 million bushels are produced and consumed. Entry of new firms reduces price and raises total output. (Had the price not fallen, quantity supplied after entry would have been 80 million bushels—point *F*.) Why must the price fall? Because the demand curve for the industry is downward sloping—consumers will purchase the increased output only if the price is reduced.

To see where the entry process stops, we must consider how the entry of new firms affects the behavior of old firms. At first glance, this notion may seem to

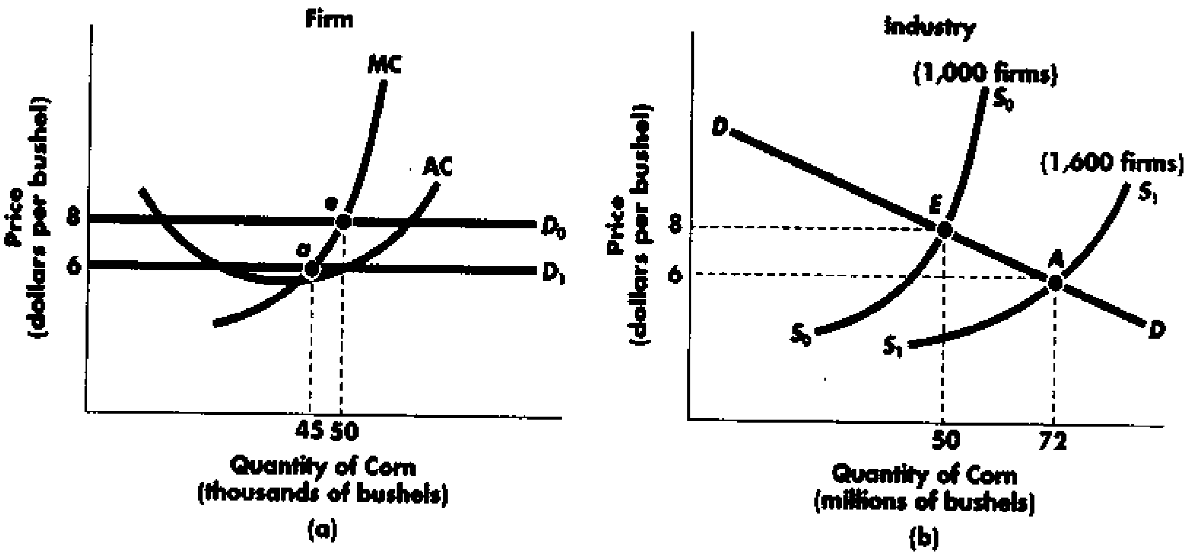
FIGURE 9-7 A SHIFT IN THE INDUSTRY SUPPLY CURVE CAUSED BY THE ENTRY OF NEW FIRMS

This diagram shows what happens to the industry equilibrium when new firms enter the industry. Quantity supplied at any given price increases; that is, the supply curve shifts to the right, from S_0S_0 to S_1S_1 in the figure. As a result, the market price falls (from $8 to $6) and the quantity supplied increases (from 50 million bushels to 72 million bushels).



contradict the idea of perfect competition; perfectly competitive firms are not supposed to care what their competitors are doing. Indeed, these corn farmers do not care. But they *do* care very much about the market price of corn and, as we have just seen, the entry of new firms into the corn-farming industry lowers the price of corn.

FIGURE 9-8 THE COMPETITIVE FIRM AND THE COMPETITIVE INDUSTRY



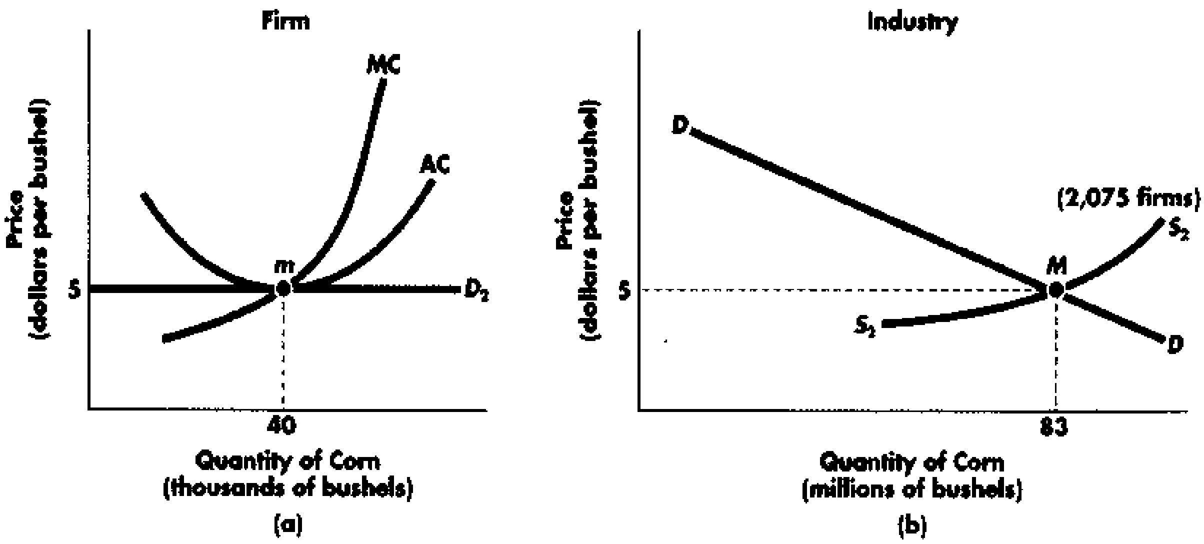
Here we show the interaction between developments at the industry level, in Panel (b), and developments at the firm level, in Panel (a). An outward shift in the industry supply curve from S_0S_0 to S_1S_1 in Panel (b) lowers the market price from $8 to $6. In Panel (a), we see that a profit-maximizing competitive firm reacts to this decline in price by curtailing output. When the demand curve of the firm is D_0 (height = $8), it produces 50,000 bushels (point *e*). When the firm's demand curve falls to D_1 (height = $6), its output declines to 45,000 bushels (point *a*). However, there are now 1,600 firms rather than 1,000, so total industry output has expanded from 50 million bushels to 72 million bushels, as in Panel (b). Entry has reduced profits. But since *P* still exceeds AC at an output of 45,000 bushels per firm in Panel (a), some profits remain.

In Figure 9-8, we juxtapose the diagram of the equilibrium of the competitive firm (Figure 9-2 on page 214) and the diagram of the equilibrium of the competitive industry (Figure 9-7). Before entry, the market price was \$8, point *E* in Figure 9-8(b), and each of the 1,000 firms was producing 50,000 bushels—the point where marginal cost and price were equal, point *e* in Figure 9-8(a). The demand curve facing each firm was the horizontal line D_0 in Figure 9-8(a). There were profits because average costs (AC) at 50,000 bushels per firm were less than price.

Now suppose that 600 new firms are attracted by these high profits and enter the industry. Each has the cost structure indicated by the AC and MC curves in Figure 9-8(a). As we have noted, the industry supply curve in Figure 9-8(b) shifts to the right, and price falls to \$6 per bushel. Firms in the industry cannot fail to notice this lower price. As we see in Figure 9-8(a), each firm reduces its output to 45,000 bushels in reaction to the lower price (point *a*). But now there are 1,600 firms, so total industry output is 45,000 × 1,600 = 72 million bushels, point *A* in Figure 9-8(b).

At point *a* in Figure 9-8(a), there are still profits to be made because the \$6 price exceeds average cost. Thus, the entry process is not yet complete. It will end only when all profits have been competed away. The two panels of Figure 9-9 show the competitive firm and the competitive industry in long-run equilibrium. Only when entry shifts the industry supply curve so far to the right, S_2S_2 in Figure 9-9(b), that the demand curve facing individual firms falls to the level of minimum average cost, point *m* in Figure 9-9(a), will all profits be eradicated and entry cease.[4]

FIGURE 9-9 LONG-RUN EQUILIBRIUM OF THE COMPETITIVE FIRM AND INDUSTRY



By the time 2,075 firms have entered the industry, the industry supply curve is S_2S_2 and the market price is \$5 per bushel. At this price, the horizontal demand curve facing each firm is D_2 in Panel (a), so the profit-maximizing level of output is 40,000 bushels (point *m*). Here, since average cost and price are equal, there is no economic profit.

[4] If the original short-run equilibrium had involved losses instead of profits, firms would have exited from the industry, shifting the industry supply curve inward, until all losses were eradicated, and we would end up in a position exactly like Figure 9-9. EXERCISE: To test your understanding, draw the version of Figure 9-8 that corresponds to this case.

Notice that at the equilibrium point, *m* in Panel (a), each firm picks its own output level so as to maximize its profit. This means that for each firm $P = MC$. But free entry forces AC to be equal to P in the long run [point M in Panel (b) of the graph], for if P were not equal to AC, firms would either earn profits or suffer losses. That would mean, in turn, that firms would find it profitable to enter or leave the industry, which is incompatible with industry equilibrium. Thus:

When a perfectly competitive industry is in long-run equilibrium, firms maximize profits so that $P = MC$, and entry forces the price down until it is tangent to the long-run average cost curve ($P = AC$). As a result, in long-run competitive equilibrium it is always true that:

$$P = MC = AC$$

Thus, even though every firm earns zero profit, profits are at the maximum that is attainable.[5]

THE LONG-RUN INDUSTRY SUPPLY CURVE

We have now seen basically what lies behind the supply-demand analysis that was first introduced in Chapter 4. Only one thing remains to be explained. Figures 9-5 through 9-8 depicted short-run industry supply curves and short-run equilibrium. However, since Figure 9-9 describes long-run competitive equilibrium, its industry supply curve must also, obviously, pertain to the long run.

How is the long-run industry supply curve related to the short-run supply curve? The answer is implicit in what we have just discussed. The long-run industry supply curve evolves from the short-run supply curve via two simultaneous processes. First, there is the entry of new firms or the exit of old ones, which shifts the short-run industry supply curve toward its long-run position.

Second, and concurrently, as each firm in the industry is freed from its sunk commitments, the cost curves pertinent to its decisions become its long-run cost curves rather than its short-run cost curves. For example, consider a company that was stuck in the short run with a plant designed to serve 20,000 customers even though it is fortunate enough to have 25,000 customers. When it is time to replace the old plant, management will obviously build a new plant that can serve the larger number of customers more conveniently and efficiently. The reduced cost that results from the larger plant is the cost that is pertinent to both the firm and the industry in the long run.

Finally, let us note that the long-run supply curve of the competitive industry (S_2S_2 in Figure 9-9) must be identical to the industry's long-run *average* cost curve. This is so because in the long run, as we have seen, economic profit must be zero. The price the industry charges cannot exceed the long-run average cost (LRAC) of supplying that quantity because any excess of price over LRAC would constitute a profit opportunity that would attract new firms. Similarly, price cannot be below LRAC because firms would then refuse to supply that output at this price. Therefore, for each possible long-run quantity supplied, the price must equal the industry's long-run average cost. It is this long-run industry supply curve that is relevant to the determination of long-run equilibrium price and quantity in a standard supply-demand diagram.

---

[5]EXERCISE: Show what happens to the equilibrium of the firm and of the industry in Figure 9-9 if there is a rise in consumer income that leads to an outward shift in the industry demand curve.

These ideas are illustrated in Figure 9-10, in which the short-run industry supply curve, *SS*, is above and to the left of the long-run average cost curve, LRAC. Consider any industry output, say, 70 million bushels of corn per year. At that output, the long-run average cost is $4 per bushel (point *A*). But if the price charged by farmers were given by the short-run supply curve for that output—that is, $7 per bushel (point *B*)—then the firms would earn $3 in economic profit on each and every bushel they sold.

Such economic profits would induce other firms to enter the industry, which would force prices downward as the industry supply curve shifted outward. So long as this shift did not take *SS* all the way down to LRAC, some economic profits would remain, and so entry would continue. Thus, *SS* must continue to fall until it reaches the position of the long-run average cost curve. Then and only then will entry cease and long-run equilibrium be attained.

The long-run supply curve of the competitive industry is also the industry's long-run average cost curve. The industry is driven to that supply curve by the entry or exit of firms and by the adjustment of firms already in the industry.

ZERO ECONOMIC PROFIT: THE OPPORTUNITY COST OF CAPITAL

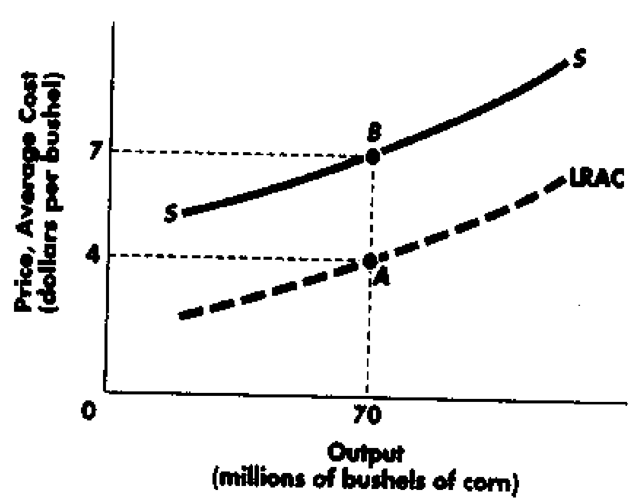
Something may be troubling to you about our discussions of the long run. Why would there be any firms in the industry *at all* if in the long run there were no profits to be made? What sense does it make to call a position of zero profit a "long-run equilibrium"? The answer is that the zero profit concept used in economics does not mean the same thing that it does in ordinary usage.

We have noted repeatedly that when economists measure average cost, they include the cost of *all* of the firm's inputs, *including the opportunity cost of the capital or any other inputs, such as labor, provided by the firm's owners*. Since the firm may not make explicit payments to some of those who provide it with capital, this element of cost may not be picked up by the firm's accountants. So what economists call *zero economic profit* will correspond to some positive amount of profit as measured by conventional accounting techniques. For example, if

FIGURE 9-10

SHORT-RUN INDUSTRY SUPPLY AND LONG-RUN INDUSTRY AVERAGE COST

If the industry supply curve (*SS*) lies above and to the left of its long-run average cost curve (LRAC), the industry must earn economic profits (average economic profit = the distance *AB*). That will induce entry, increase output at any given price, and shift the supply curve to the right until it coincides with the long-run average cost curve.



investors can earn 15 percent by lending their funds elsewhere, then the firm must earn a 15 percent rate of return to cover its opportunity cost of capital.

Because economists consider this 15-percent opportunity cost to be the *cost of the firm's capital*, they include it in the AC curve. If the firm cannot earn at least 15 percent on its capital, funds will not be made available to it, because investors can earn greater returns elsewhere. So, in the economist's language, in order to break even—earn zero **economic profit**—a firm must earn enough not only to cover the cost of labor, fuel, and raw materials, but also the cost of its funds, including the opportunity cost of any funds supplied by the owners of the firm.

ECONOMIC PROFIT equals net earnings, in the accountant's sense, minus the opportunity costs of capital and of any other inputs supplied by the firm's owners.

To illustrate the difference between economic profit and accounting profit, suppose that U.S. government bonds pay 8 percent, and the owner of a small shop earns 6 percent on her business investment. The shopkeeper might say that she is making a 6 percent profit, but an economist would say that she is *losing* 2 percent on every dollar she has invested in her business. The reason is that by keeping her money tied up in the firm, she gives up the chance to buy government bonds and receive an 8 percent return. With this explanation of the meaning of economic profit, we can now understand the logic behind the zero-profit condition for the long-run industry equilibrium.

Zero profit in the economic sense simply means that firms are earning the normal, economy-wide rate of profit in the accounting sense. This result is guaranteed, in the long run, under perfect competition by freedom of entry and exit.

Freedom of entry guarantees that those who invest in a competitive industry will receive a rate of return on their capital *no greater than* the return that capital could earn elsewhere in the economy. If economic profits were being earned in some industry, capital would be attracted into it. The new capital would shift the industry supply curve to the right, which would drive down prices and profits. This process would continue until the return on capital in this industry was reduced to the return that capital could earn elsewhere—its opportunity cost.

Similarly, freedom of exit of capital guarantees that, in the long run, once capital has had a chance to move, no industry will provide a rate of return *lower than* the opportunity cost of capital. For if returns in one industry were particularly low, resources would flow out of it. Plant and equipment would not be replaced as it wore out. As a result, the industry supply curve would shift to the left, and prices and profits would rise toward their opportunity cost level.

PERFECT COMPETITION AND ECONOMIC EFFICIENCY

Economists have long admired perfect competition as a thing of beauty, like one of King Tutankhamen's funerary masks. (It's just as rare!) Adam Smith's invisible hand produces results that are considered *efficient* in a variety of senses that we will examine carefully in the next chapter. But one aspect of the great efficiency of perfect competition follows immediately from the analysis we have just completed.

We have seen earlier that, when the firm is in long-run equilibrium, it must have P = MC = AC, as indicated by Figure 9-9(a) on page 223. This implies that the long-run competitive equilibrium of the firm will occur at the lowest point on its long-run AC curve, which is also where that curve is tangent to the firm's horizontal demand curve.

In long-run competitive equilibrium, every firm produces at the minimum point on its average cost curve. Thus, the outputs of competitive industries are produced at the lowest possible cost to society.

POLICY DEBATE — SHOULD PERFECT COMPETITION BE USED AS A GUIDE BY GOVERNMENT REGULATORS?

As we have seen here, and will discuss further in the following chapter, perfect competition displays the market mechanism at its best. It prevents firms from earning excess profits. It forces firms to produce the output quantity at which AC is as low as possible, and it has other virtues besides.

As we will see in Chapters 11 and 12, markets where monopoly or oligopoly prevail are very different from perfect competition. In monopolistic or oligopolistic markets a few large firms may charge high prices that yield large profits, and they may produce output quantities that do not match consumer preferences. Consequently (see Chapter 18), such industries are often regulated by government agencies.

But what should regulation seek to force monopoly or oligopoly firms to do? Should it force them to behave like perfectly competitive firms? Should it force their prices to equal marginal costs? Should it try to break them up into thousands of tiny enterprises?

No one believes that government regulation should go quite so far. Indeed, there are economists and others who argue that perfect competition is an undesirable and, indeed, an impossible goal for such industries. For example, if those industries have economies of scale, breaking them into small firms will raise their costs and consumers will have to pay more, not less. Moreover, as we saw in Chapter 7, where there are economies of scale, the average cost curve must go downhill—the larger the firm's output, the lower its average cost. So marginal cost must be below average cost (see the appendix to Chapter 7 for review), and a price equal to marginal cost must also be below average cost. Thus, where there are scale economies, if the firm is forced to charge a price equal to marginal cost it will be forced to go bankrupt!

Still, there are regulators, economists, and others who believe that perfect competition is so desirable a state of affairs that regulated firms should be required to come as close to it as possible in their behavior.

An example will show why it is most efficient if each firm in a competitive industry produces at the point where AC is as small as possible. Suppose the industry is producing 12 million bushels of corn. This amount can be produced by 120 farms each producing 100,000 bushels, by 100 farms each producing 120,000 bushels, or by 200 farms each producing 60,000 bushels. Of course, the job can also be done instead by other numbers of farms, but for simplicity let us consider only these three possibilities.

Suppose that the AC figures for the firm are as shown in Table 9-3. Suppose, moreover, that an output of 100,000 bushels corresponds to the lowest point on the AC curve, with an AC of 70 cents per bushel. Which is the cheapest way for the industry to produce its 12-million-bushel output? That is, what is the cost-minimizing number of firms for the job? Looking at the last column of Table 9-3, we see that the industry's total cost of producing the 12-million-bushel output is as low as possible if it is done by 120 firms each producing the cost-minimizing output of 100,000 bushels.

TABLE 9-3

AVERAGE COST FOR THE FIRM AND TOTAL COST FOR THE INDUSTRY

Firm's Output (bushels)	Firm's Average Cost (dollars)	Number of Firms	Industry Output (bushels)	Total Industry Cost (dollars)
60,000	0.90	200	12,000,000	$10,800,000
100,000	0.70	120	12,000,000	8,400,000
120,000	0.80	100	12,000,000	9,600,000

Why is this so? The answer is not difficult to see. For any given industry output, Q, it is clear that *total* industry cost (= AC × Q) will be as small as possible if and only if AC for *each* firm is as small as possible, that is, if the number of firms doing the job is such that each is producing the output at which AC is as low as possible.

That this kind of cost efficiency characterizes perfect competition in the long run can be seen in Figures 9-8 (page 222) and 9-9 (page 223). Before full long-run equilibrium is reached (Figure 9-8), firms may not be producing in the least costly way. For example, the 50 million bushels being produced by 1,000 firms at points e and E in Figures 9-8(a) and 9-8(b) could be produced more cheaply by more firms, each producing a smaller volume, because the point of minimum average cost lies to the left of point e in Figure 9-8(a). This problem is rectified, however, in the long run by entry of new firms seeking profit. We see in Figure 9-9 that after the entry process is complete, every firm is producing at its most efficient (lowest AC) level—40,000 bushels.

As Adam Smith might have put it, even though each farmer cares only about his or her own profits, the corn-farming industry as a whole is guided *by an invisible hand* to produce the amount of corn that society wants at the lowest possible cost.

OUR TWO PUZZLES RESOLVED

We end the chapter by returning to the two puzzles with which it began, because we now have all the tools needed to resolve them. We begin with the alternative pollution-curbing proposals.

ISSUE 1: CUTTING POLLUTION—THE CARROT OR THE STICK?

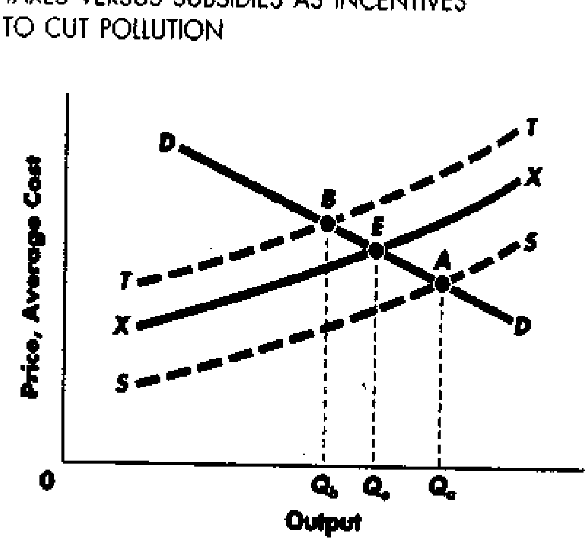
Recall that we asked: Should polluters be taxed on their emissions, or should they, instead, be offered *subsidies* to cut emissions? A subsidy would indeed induce firms to cut their emissions. Nevertheless, the paradoxical result is likely to be an increase in total pollution. Let us see now why this is so.

In Figure 9-11, we depict the long-run average cost curve (LRAC) of the industry, XX. We now know that this must also be the industry's long-run supply curve, because if the supply curve lies above (to the left of) LRAC, economic profits would be earned and entry would drive the supply curve to the right. The opposite would occur if the supply was below and to the right of LRAC.

Now, a tax on business firms clearly raises the long-run average costs of the industry. Suppose that it shifts the LRAC, and thus the long-run supply curve upward from XX to TT in the graph. This would move the equilibrium point from E to B. Similarly, a subsidy reduces average cost, and so it shifts the LRAC and the long-run supply curve downward and to the right (from XX to SS). This would move the equilibrium point from E to A.

Our paradoxical result follows from the presumption that the more output a polluting industry produces, the more pollution it will emit. Under the tax on emissions, equilibrium moves from E to B, and so output falls from Q_e to Q_b. Thus, emissions will fall—just as common sense leads us to expect. But, with the subsidy, industry output would *rise* from Q_e to Q_a. Thus, contrary to intuition and despite the fact that each firm emits less, the industry must pollute more!

FIGURE 9-11 TAXES VERSUS SUBSIDIES AS INCENTIVES TO CUT POLLUTION



Because a tax on emissions raises industry average cost, it will also raise the industry supply curve (the shift from *XX* to *TT*). This will move the equilibrium to point *B* and reduce output from Q_e to Q_b. The result is a cut in pollution. A subsidy for pollution reductions, however, reduces average cost and so shifts the industry supply curve downward to *SS*. The equilibrium point then moves to *A* and raises output to Q_a. Thus, the subsidy backfires by attracting so many additional polluting firms into the industry.

What explains this paradox? The answer is *entry*. The subsidy will initially bring economic profits to the polluters, and that will attract even more polluters into the industry. In essence, a subsidy encourages more polluters to open up for business. But our graph takes us one step beyond this simple observation. It is true that we end up with more polluting firms, but each will be polluting less than before. Thus, we have one influence leading to more pollution and another influence leading to less pollution. Which of these forces will win out? The graph tells us that if a rise in the polluting good's output always increases pollution, then subsidy *must* lead to increased pollution on balance.

ISSUE 2: CAN FIRMS SHIFT A TAX TO THEIR CONSUMERS?

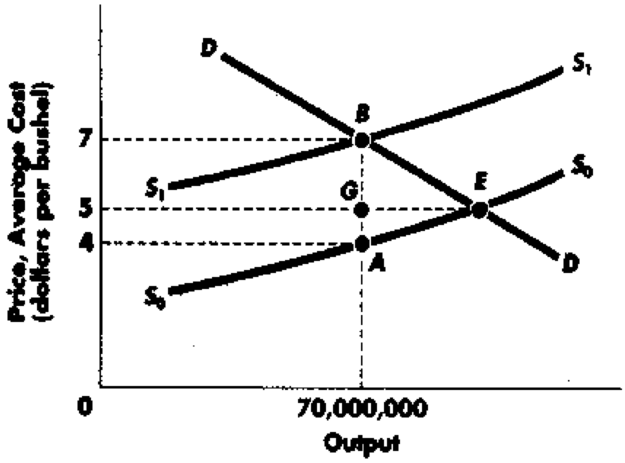
Our second puzzle was why firms so strongly resist taxes on their products if they can simply raise prices and make consumers bear the costs. As we will see now, a tax of, say, $3 per bushel of corn will generally raise the market price of corn by a smaller amount, say, $2. If that is so, then only $2 of the $3 tax will have been shifted to consumers. The remaining $1 will somehow have to be borne by the suppliers. Let us see how.

In Figure 9-12, we assume that a $3 tax is levied on every bushel of corn sold. This means that average cost must rise by exactly $3 at every output level. For example, we see in the graph that if output is 70 million bushels per year, then average cost rises from $4 before the tax (point *A*) to $7 after the tax is imposed (point *B*). Reasoning in this way, we see that the long-run industry supply curve must shift upward by exactly $3 at every output level. Thus, the industry supply curve shifts up from S_0S_0 to S_1S_1.

As a result, we see that the equilibrium point moves upward and to the left, from *E* to *B*, raising the equilibrium price from $5 to $7. This $2 price increase is only two-thirds of the $3 tax. In general terms, the price rise is shown by the distance *GB*, while the tax is the larger distance, *AB*. The burden is therefore shared between consumers and producers.

How are consumers able to protect themselves from being stuck with the entire tax? The answer is simple. When price rises, consumers have a way of punishing suppliers: they simply purchase less. That is the meaning of the

FIGURE 9-12 CAN FIRMS SHIFT A TAX TO THEIR CONSUMERS?



A \$3 tax per unit of output raises average cost by \$3, so the entire long-run supply curve shifts up by that amount (from S_0S_0 to S_1S_1). This moves the equilibrium point from *E* to *B* and raises price from \$5 to \$7 (from *G* to *B*). Thus, the price rise pays for only part of the \$3 tax (distance *AB*). Consumers do not bear the entire burden of the tax because of the decrease in quantity demanded in response to any rise in price—the negative slope of the demand curve.

negative slope of the demand curve—the higher the price, the less consumers will buy. In fact, it can be shown that the flatter the demand curve (that is, the more elastic it is), the less of any given tax consumers will have to bear. (See Review Question 11 at the end of the chapter and a further discussion in Chapter 20.)

Does all this mean that profits will fall because of the tax? The surprising answer is no—at least not in the long run. Remember that long-run equilibrium profit is always zero under perfect competition, with or without the tax. So the tax cannot change the long-run profit rate. Then why does industry object to such a tax? Because, with a negatively sloped demand curve, a tax will cut into sales, cause short-term losses, and force some firms out of the industry. That is not a pleasant prospect for incumbent firms.

SUMMARY

1. **Markets** are classified into several types depending on the number of firms in the industry, the degree of similarity of their products, and the possibility of impediments to entry.
2. The four main market structures discussed by economists are monopoly (single-firm production), oligopoly (production by a few firms), monopolistic competition (production by many firms with somewhat different products), and **perfect competition** (production by many firms with identical products and free entry and exit).
3. Few, if any, industries satisfy the conditions of perfect competition exactly, although some come close. Perfect competition is studied because it is easy to analyze and because it is useful as a yardstick to measure the performance of other market forms.
4. The demand curve of the perfectly competitive firm is horizontal because its output is so small a share of the industry's production that it cannot affect price. With a horizontal demand curve, price, average revenue, and marginal revenue are all equal.
5. The short-run equilibrium of the perfectly competitive firm is at the level of output that maximizes profits; that is, where MR = MC = price. This equilibrium may involve either a profit or a loss.
6. The short-run supply curve of the perfectly competitive firm is the portion of its marginal cost curve that lies above its average variable cost curve.
7. The industry's short-run supply curve under perfect competition is the horizontal sum of the supply curves of all of its firms.
8. In the long-run equilibrium of the perfectly competitive industry, freedom of entry forces each firm to earn zero **economic profit**, that is, no more than the firm's capital could earn elsewhere (the opportunity cost of the capital).

9. Industry equilibrium under perfect competition is at the point where the industry supply and demand curves intersect.
10. In long-run equilibrium under perfect competition, the firm chooses output such that average cost, marginal cost, and price are all equal. Output is at the point of minimum average cost. The firm's demand curve is tangent to its average cost curve at its minimum point.
11. The competitive industry's long-run supply curve coincides with its long-run average cost curve.
12. Both a tax on emission of pollutants and a subsidy payment for reductions in emissions induce firms to cut emissions. However, under perfect competition, a subsidy leads to the entry of more polluting firms and the likelihood of a net increase in emissions by the industry.
13. A tax on a good generally leads to a rise in the price at which the taxed product is sold. However, the rise in price is generally less than the tax, so consumers usually pay less than the entire tax.
14. Consumers generally pay only part of a tax because the resulting rise in price leads them to buy less and the cut in the quantity they demand helps to keep price down.

KEY TERMS

Market
Perfect competition
Pure monopoly
Monopolistic competition
Oligopoly
Price taker
Horizontal demand curve
Short-run equilibrium
Sunk cost
Variable cost
Supply curve of the firm
Supply curve of the industry
Long-term equilibrium
Opportunity cost
Economic profit

QUESTIONS FOR REVIEW

1. Explain why a perfectly competitive firm does not expand its sales without limit if its horizontal demand curve means that it can sell as much as it wants to at the current market price.
2. Explain why a demand curve is also a curve of average revenue. Recalling that when an average revenue curve is neither rising nor falling, marginal revenue must equal average revenue, explain why it is always true that $P = MR = AR$ for the perfectly competitive firm.
3. Under what circumstances might you expect the demand curve of the firm to be (a) vertical? (b) horizontal? (c) negatively sloping? (d) positively sloping?
4. Explain why $P = MC$ in the short-run equilibrium of the perfectly competitive firm, while in long-run equilibrium $P = MC = AC$.

5a. Which of the four attributes of perfect competition (many small firms, freedom of entry, standardized product, perfect information) is primarily responsible for the fact that the demand curve of a perfectly competitive firm is horizontal?

5b. Which of the four attributes of perfect competition is primarily responsible for the firm's zero economic profits in long-run equilibrium?

6. We indicated in the chapter (page 218) that the MC curve cuts the AVC (average variable cost) curve at the *minimum* point of the latter. Explain why this must be so. (*Hint:* Since marginal costs are, by definition, entirely composed of variable costs, the MC curve can be considered the curve of *marginal variable costs*. Apply the general relationships between marginals and averages explained in Chapter 8.)
7. Explain why it is not sensible to close a business firm if it earns zero economic profits.
8. If the firm's lowest average cost is \$48 and the corresponding average variable cost is \$24, what does it pay a perfectly competitive firm to do if:
   a. the market price is \$49?
   b. the price is \$30?
   c. the price is \$8?
9. If the market price in a competitive industry were above its equilibrium level, what would you expect to happen?
10. Draw a long-run supply curve for a competitive industry. Assume that the government imposes a tax of \$2 per unit of the product sold. What will happen to the supply curve? Why?
11. Draw two graphs like Figure 9-12. In one of them, make the demand curve fairly flat; in the other, make

it fairly steep. Which of the demand curves is more elastic? In which of them is a larger share of the tax shifted to consumers? Why?

12. (More difficult) In this chapter we stated that the firm's MC curve goes through the lowest point of its AC curve and also through the lowest point of its AVC curve. Since the AVC curve lies below the AC curve, how can both of these statements be true? Why are they true? (*Hint:* see Figure 9-4.)

CHAPTER 10

THE PRICE SYSTEM AND THE CASE FOR LAISSEZ-FAIRE

If there existed the universal mind that . . . would register simultaneously all the processes of nature and of society, that could forecast the results of their inter-reactions, such a mind . . . could . . . draw up a faultless and an exhaustive economic plan. . . . In truth, the bureaucracy often conceives that just such a mind is at its disposal; that is why it so easily frees itself from the control of the market.

Leon Trotsky (a leader of the Russian Revolution)

Our study of microeconomics is focused on a crucial question: What does the market do well, and what does it do poorly? By applying what we learned about demand in Chapters 5 and 6 and supply in Chapters 7 and 8, we can provide a fairly comprehensive answer to the first part of this question. This chapter will describe the tasks that the market carries out well—some, indeed, with spectacular effectiveness.

We begin by recalling two important themes of Chapters 3 and 4. First, because all resources are scarce, it is critical to utilize them *efficiently;* second, an economy must have some way to *coordinate* the actions of many individual consumers and producers. Specifically, society must somehow choose *how much* of each good to produce, *what input quantities* to use in the production process, and *how to distribute* the resulting outputs among consumers. As suggested by the opening quotation (by an author who was in a position to know), these tasks are exceedingly difficult for central planners, but they are rather simple for a market system. This is why observers with philosophies as diverse as those of Adam Smith and Leon Trotsky have been admirers of the market, and why more and more formerly Marxist countries have now moved toward market economies.

But the chapter should not be misinterpreted as a piece of salesmanship, for that is not its purpose. Instead, we shall study an idealized price system in which every good is produced under the exacting conditions of perfect competition. While a number of industries in our economy are reasonable approximations to perfect competition, others are as different from this idealized world as the physical world is from a frictionless vacuum tube. But just as the physicist uses the vacuum tube to study the laws of gravity, the economist uses the theoretical concept of a perfectly competitive economy to analyze the virtues of the market. There will be plenty of time in later chapters to study its vices.

PUZZLE: SAN FRANCISCO BAY BRIDGE PRICING

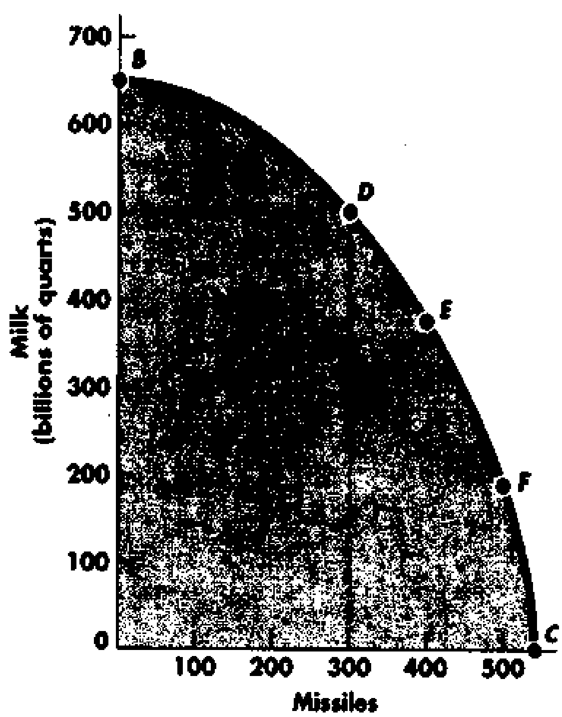
Appropriate pricing is the key to efficient resource allocation in a market economy. In California, the San Francisco–Oakland Bay Bridge is very heavily used. The large volume of toll-paying traffic has probably long since paid for the cost of building this bridge, while that is probably less likely in the cases of the nearby San Mateo–Hayward and Dumbarton bridges, which are less crowded. Yet economists argue that the price charged to use the San Francisco–Oakland Bay Bridge should be *higher* than the prices for the second two. Why might that make sense? Before you have finished reading this chapter you will be able to answer that question.

EFFICIENT RESOURCE ALLOCATION: THE CONCEPT

The fundamental fact of scarcity limits the volume of goods and services that any economic system can produce. In Chapter 3, we illustrated the concept of scarcity with a graphic device called a *production possibilities frontier,* which we repeat here for convenience as Figure 10-1. The frontier, curve *BC*, depicts all combinations of missiles and milk that this society can produce given the limited resources at its disposal. For example, if it decides to produce 300 missiles, it will have enough resources left over to produce *no more than* 500 billion quarts of milk (point *D*). Of course, it is always possible to produce fewer than 500 billion quarts of milk—at a point, such as *G*, below the production possibilities frontier. But if a society does this, it is wasting some of its productive potential; that is, it is not operating *efficiently*.

FIGURE 10-1 PRODUCTION POSSIBILITIES FRONTIER AND EFFICIENCY

Every point on the production possibilities frontier, *BC*, represents an efficient allocation of resources because it is impossible to get more of one item without giving up some of the other. Any point below the frontier, like *G*, is inefficient, since it wastes the opportunity to obtain more of both goods.



An **EFFICIENT ALLOCATION OF RESOURCES** is one that takes advantage of every opportunity to make some individuals better off in their own estimation while not worsening the lot of anyone else.

In Chapter 3, we defined *efficiency* rather loosely as the absence of waste. Since the main subject of this chapter is how a competitive market economy allocates resources efficiently, we now need a more precise definition. It is easiest to define an **efficient allocation of resources** by saying what it is *not*. For example, suppose that we could rearrange our allocation of resources so that one group of people would get more of the things it wanted, while at the same time no one else would have to give up anything. Then, the failure to change the allocation of resources to take advantage of this opportunity would surely be wasteful—that is, *inefficient*. When society has taken advantage of every such opportunity for improvement, so that no such possibilities remain for making some people better off without making others worse off, we say that the allocation of resources is *efficient*.

Because point *G* in Figure 10-1 is below the frontier, there *must* be points like *E* on the frontier that lie above and to the right of *G*. This means that at *E* we get more of *both* outputs without any increase in input! Thus, *no point below the frontier* can represent an efficient allocation of resources. By contrast, *every point on the frontier* is efficient because, no matter where on the frontier we start, it is impossible to get more of one good without giving up some of the other.

This discussion also shows that there are normally many efficient allocations of resources; in the example, *every* point on frontier *BC* can be efficient. As a rule, the concept of efficiency does not permit us to tell which allocation is "best" for society. Yet, as we shall see in this chapter, the concept of efficiency can be used to formulate surprisingly detailed rules to steer us away from situations in which resources are being wasted.

PRICING TO PROMOTE EFFICIENCY: AN EXAMPLE

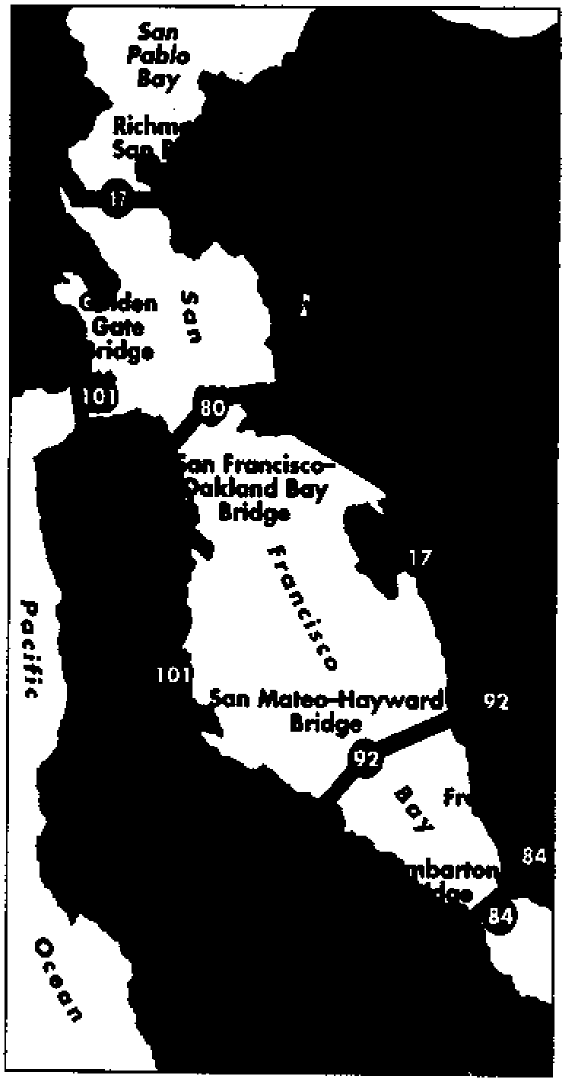
We can illustrate the meaning of *efficiency* and how the choice of prices can make the difference between efficiency and inefficiency by using the real example in our opening puzzle—the prices (tolls) that are charged to use the bridges in the San Francisco Bay area. We will see that, although proper pricing of these bridges can enhance efficiency, people nonetheless may well reject the efficient solution.

Figure 10-2 shows a map of the San Francisco Bay area, featuring the five bridges that serve the bulk of the traffic in and around the bay. A traveler going from north of Berkeley (point *A*) to Palo Alto (point *B*) has a choice of at least three routes:

1. Over the Richmond–San Rafael Bridge, across the Golden Gate Bridge, through San Francisco, and on southward.
2. Across the bay on the San Francisco–Oakland Bay Bridge and on southward as before.
3. Down the eastern side of the bay, across the San Mateo–Hayward Bridge or the Dumbarton Bridge, and then on to Palo Alto (this route is in orange in Figure 10-2).

Let's consider which of these three choices utilizes society's resources most efficiently. The most crowded of the five bridges is the San Francisco–Oakland Bay Bridge, followed by the Golden Gate Bridge. The first carries 254,000 vehicles per day, and the second 116,000. During rush hours (when, for instance, over 20,000 vehicles per hour use the San Francisco–Oakland Bay Bridge), delays are frequent and traffic barely crawls across both bridges. In other words, space is

FIGURE 10-2 TOLL BRIDGES OF THE SAN FRANCISCO BAY AREA



scarce, and every car that uses these bridges makes it that much harder for others to get across.

On the other hand, the San Mateo–Hayward and Dumbarton bridges carry approximately 72,000 and 52,000 vehicles per day, respectively. To achieve efficiency, any driver who is indifferent about the two routes should take the one using the least crowded bridges. This would help reduce the amount of travel time wasted by the population as a whole. Specifically, in our illustration, Route 1, using the Golden Gate Bridge, is not a socially desirable way for our driver to get to Palo Alto. Route 2, with its use of the San Francisco–Oakland Bay Bridge, is even worse because of the added delays it causes for everyone else. Route 3, for drivers who are indifferent about these options, is the best choice from the viewpoint of the public interest. This is not meant to imply that it is socially efficient to equalize the traffic among the routes, but it certainly will help travelers get where they are going faster if some of the traffic can be induced to leave the most crowded routes and switch over to some less crowded ones.

EARTHQUAKE, BRIDGE CONGESTION, AND ROUTE SUBSTITUTION

Bridge congestion and route substitution took on graphic proportions after the partial collapse of the San Francisco–Oakland Bay Bridge during the Loma Prieta earthquake that shook the San Francisco area on October 17, 1989. According to one newspaper report:

> The bridge closure forced commuters ... [who] used it to cross the bay some other way. While many people stayed home the week after the disaster, they now appear to be rejoining the ranks of the daily rush.... The result has been major delay on alternative routes, such as the Golden Gate, San Mateo and San Rafael–Richmond bridges.

Just a little over four years later the Northridge earthquake in Los Angeles decimated that city's massive freeway system, with the most dramatic damage to bridges on the nation's most heavily traveled road, the Santa Monica Freeway. Many of the more than 300,000 commuters who travel that highway daily switched over to the city's rail system, where average ridership shot up from 8,000 commuters per day to peaks of 30,000 per day.

SOURCES: *The Boston Globe*, October 31, 1989, p. 15; "New Technology Cuts Time for Bridge Repair," *American City and County*, December 1994, p. 42; "Working Together to Unclog Road," *Public Relations Journal*, May 1994, pp. 7–8.



Appropriate prices can promote efficiency in the utilization of the bridges. Specifically, if we were to charge higher prices (very likely substantially higher prices) for the use of the most crowded bridges, on which space is a scarce resource, balanced by lower prices on the uncrowded bridges, we could induce still more drivers to use the uncrowded bridges. This is just the same reasoning that leads economists to advocate low prices for abundant minerals and high prices for scarce ones.

CAN PRICE INCREASES EVER SERVE THE PUBLIC INTEREST?

This last discussion raises a point that people untrained in economics always find difficult to accept: *low prices may not always be in the public interest.* The reason is clear enough. If a price, such as the price of crossing a crowded bridge or the price of oil, is set "too low," then consumers will receive the "wrong" signals. They will be encouraged to crowd the bridge even more or to consume more oil, thereby squandering society's precious resources.

A historic illustration is perhaps the most striking way to bring out the point. In 1834, some 10 years before the great potato famine caused unspeakable misery and death by starvation and brought so many people from Ireland to the United States, a professor of economics named Mountifort Longfield lectured at

the University of Dublin about the price system. He offered the following remarkable illustration of his point:

> Suppose the crop of potatoes in Ireland was to fall short in some year one-sixth of the usual consumption. If [there were no] increase of price, the whole ... supply of the year would be exhausted in 10 months, and for the remaining 2 months a scene of misery and famine beyond description would ensue.... But when prices [increase] the sufferers [often believe] that it is not caused by scarcity.... They suppose that there are provisions enough, but that the distress is caused by the insatiable rapacity of the possessors ... [and] they have generally succeeded in obtaining laws against [the price increases] ... which alone can prevent the provisions from being entirely consumed long before a new supply can be obtained.[1]

Longfield's reasoning can usefully be rephrased. If the crop fails, potatoes become scarcer. If society is to use its scarce resources efficiently, stretching out the potato supply to last until the next crop arrives, it must cut back on the consumption of potatoes during earlier months—which is just what rising prices will do *automatically* if free-market mechanisms are allowed to work. However, if the price is held artificially low, consumers will use society's resources inefficiently. In this case, the inefficiency shows up in the form of famine and suffering when the year's crop is consumed months before the next one is harvested.

It is not easy to accept the notion that higher prices can serve the public interest better than lower ones. Politicians who voice this view are put in the position of the proverbial father who, before spanking his child, announces, "This is going to hurt me much more than it hurts you!" Since advocacy of higher prices courts political disaster, the political system often rejects the market solution when resources suddenly become more scarce.

The pricing of landings at crowded airports offers a good example. Airports are particularly congested at the "peak hours," just before 9 A.M. and just after 5 P.M., and that is when passengers most often suffer long delays. But many airports continue to charge bargain landing fees throughout the day, even at those crowded hours. That makes it attractive even for small corporate jets or other planes carrying only a few passengers to arrive and take off at those hours, worsening the delays. Higher fees for peak-hour landings can discourage such overuse, but they are politically unpopular, and many of the airports are run by local governments. Therefore, the ordinary passenger continues to experience late arrivals as a normal feature of air travel.

Keeping prices low where a rise is appropriate can have serious consequences indeed. We have seen from Longfield's example that it can contribute to famine. We know that it caused nationwide chaos in gasoline distribution after the sudden fall in Iranian oil exports in 1979. It has contributed to the surrender of cities under military siege when effective price ceilings deterred those who would otherwise have risked smuggling food supplies through enemy lines. It also has discouraged the construction of housing in cities, when rent controls made building a losing proposition.

Of course, there are cases in which it is appropriate to resist price increases—where unrestrained monopoly would otherwise succeed in gouging the public, where taxes are imposed on products capriciously and inappropriately, and where rising prices fall so heavily on the poor that rationing becomes the more acceptable option. But it is important to recognize that artificial restrictions on

[1]Mountifort Longfield, *Lectures on Political Economy Delivered in Trinity and Michaelmas Terms* (Dublin: W. Curry, Jr. and Company, 1834), pp. 53–56.

INTERFERENCES WITH THE "LAW" OF SUPPLY AND DEMAND

Recall from Chapter 4 one of the **Ideas for Beyond the Final Exam:** Interfering with free markets by preventing price increases can sometimes serve the public very badly. In extreme cases, it can even produce havoc—undermining production and causing extreme shortages of vitally needed products. The reason is that prohibiting price increases in situations of true scarcity prevents the market mechanism from reallocating resources to help cut down the shortage efficiently. The invisible hand is not permitted to do its work.

prices can produce serious and even tragic consequences—consequences that should be taken into account before a decision is made to tamper with the market mechanism.

SCARCITY AND THE NEED TO COORDINATE ECONOMIC DECISIONS

Efficiency becomes a particularly critical issue when we concern ourselves with the workings of the economy as a whole rather than narrower topics such as choosing among bridge routes or deciding on the output of a single firm. An economy may be thought of as a complex machine with literally millions of component parts. If this machine is to function efficiently, some way must be found to make the parts work in harmony.

A consumer in Peoria may decide to purchase two dozen eggs, and on the same day similar decisions are made by thousands of shoppers throughout the country. None of these purchasers knows or cares about the decisions of the others. Yet, scarcity requires that these demands must somehow be coordinated with the production process so that the total quantity of eggs demanded does not exceed the total quantity supplied. The supermarkets, wholesalers, shippers, and chicken farmers must somehow arrive at consistent decisions, for otherwise the economic process will deteriorate into chaos. And similarly with millions of other such decisions. A machine cannot run with a few missing parts.

In an economy that is planned and centrally directed, it is easy to imagine how such coordination takes place—though the implementation is far more difficult than the idea. Central planners set production targets for firms and sometimes tell firms how to meet these targets. In extreme cases, consumers may even be told, rather than asked, what they want to consume.

In a market system, prices are used to coordinate economic activity, instead. High prices discourage consumption of the resources that are most scarce, while low prices encourage consumption of the resources that are comparatively abundant. In this way, prices are the instrument used by Adam Smith's invisible hand to organize the economy's production.

The invisible hand has an astonishing capacity to handle enormously complex coordination problems—even those that remain beyond the capabilities of computers. Like any mechanism, this one has its imperfections, some of them rather serious. But we should not lose sight of the tremendously demanding task that the market constantly does accomplish—unnoticed, undirected, and in some respects, amazingly well. Let us, then, examine how the market goes about coordinating economic activity.

ECONOMIC SHOCK THERAPY IN POLAND



Since 1989, we all have witnessed the cataclysmic events in eastern Europe, including the collapse of communist central planning and the embrace of the free market. Nowhere have these changes happened as quickly and dramatically as in Poland, where reforms have constituted no less than economic "shock therapy."

The transformation of the Polish economy, though far from complete, has been nearly as radical as the first post-communist government hoped. Today, better than half of GDP comes from the private sector, up from just 29 percent in 1989. Sixty percent of the work force is now privately employed, and the annual growth rate is a healthy 5 percent. Exports, once a negligible statistic, are booming.

The new, free-market mechanisms are clearly visible on Warsaw streets. Once empty shelves are now overflowing with goods. Stores sell the latest fashions from Levi Strauss and Benetton. Glitzy billboards symbolize the battle over the consumer's pocketbook.

Despite all the good news, Poland still has its work cut out for it. Many Poles, long steeped in the constraints of a command economy, have been reluctant to embrace privatization. Unemployment has risen as useless jobs were cut. Also, transfer of state-owned companies to private hands has been a hard sell to a people who did not know what it meant to be stockholder.

SOURCES: "Tired of Capitalism? So Soon?" *The Economist*, January 1, 1995, p. 61; "Not There Yet," *The Economist*, September 3, 1994, p. 52; "Poland's Economic Reforms, If It Works, You've Fixed It," *The Economist*, January 23, 1993, p. 21; "A Survey of Central Europe, *The Economist*, November 18, 1995, p. S11.

THREE COORDINATION TASKS IN THE ECONOMY

We noted in Chapter 3 that any economic system, whether planned or unplanned, must find answers to three basic questions of resource allocation:

1. *Output selection.* How much of each commodity should be produced?
2. *Production planning.* What quantities of each of the available inputs should be used to produce each good?
3. *Distribution.* How should the resulting products be divided among consumers?

These coordination tasks may at first appear to be tailor-made for a regime of government planning like the one that used to be employed in the former Soviet Union. Yet most economists (even, nowadays, those in the formerly centrally planned economies) believe that it is in these tasks that central direction performs most poorly and, paradoxically, the undisciplined free market performs best.

To understand how the unguided market manages the miracle of creating order of what might otherwise be chaos, let us look at how each of these questions is answered by a system of free and unfettered markets—the method of economic organization that 18th-century French economists named **laissez-faire.** Under laissez-faire, the government would prevent crime, enforce contracts, and build roads and other types of public works; but it would not set prices and would interfere as little as possible with the operation of free markets. How does such an unmanaged economy solve the three coordination problems?

LAISSEZ-FAIRE refers to a program of minimal government interference with the workings of the market system. The term means that people should be left alone in carrying out their economic affairs.

OUTPUT SELECTION

A free-market system decides what should be produced via what we have called the "law" of supply and demand. Where there is a *shortage*—that is, where quan-

tity demanded exceeds quantity supplied—the market mechanism pushes the price upward, thereby encouraging more production and less consumption of the commodity that is in short supply. Where there is a *surplus*—that is, where quantity supplied exceeds quantity demanded—the same mechanism works in reverse: the price falls, discouraging production and stimulating consumption.

As an example, suppose that millions of people wake up one morning with a craving for omelets. For the moment, the quantity of eggs demanded exceeds the quantity supplied. But, within a few days, the market mechanism swings into action to meet this sudden change in demand. The price of eggs rises, which stimulates the production of eggs. In the first instance, farmers simply bring more eggs to market by taking them out of storage. Over a somewhat longer period of time, chickens that otherwise would have been sold for meat are kept in the chicken coops laying eggs. Finally, if the high price of eggs persists, farmers begin to increase their flocks, build more cages, and so on. Thus, a shift in consumer demand leads to a shift in society's resources; more eggs are wanted, and so the market mechanism sees to it that more of society's resources are devoted to the production of eggs.

Similar reactions follow if a technological breakthrough reduces the input quantities needed to produce some item. Electronic calculators are a marvelous example. Just 25 years ago, calculators were so expensive that they could be found only in business firms and scientific laboratories. Then advances in science and engineering reduced their cost dramatically, and the market went to work. With costs sharply reduced, prices fell dramatically and the quantity demanded skyrocketed. Electronics firms flocked into the industry to meet this demand, which is to say that more of society's resources were devoted to producing the calculators that were suddenly in such great demand. These examples lead us to conclude that:

Under laissez-faire, the allocation of society's resources among different products depends on two basic influences: consumer preferences and the relative difficulty of producing the goods, that is, their production costs. Prices vary so as to bring the quantity of each commodity produced into line with the quantity demanded.

Notice that no bureaucrat or central planner arranges the allocation of resources. Instead, allocation is guided by an unseen force—the lure of profits, which is the invisible hand that guides chicken farmers to increase their flocks when eggs are in greater demand and guides electronics firms to build new factories when the cost of electronic products falls.

PRODUCTION PLANNING

Once the composition of output has been decided, the next coordination task is to determine just how those goods are going to be produced. The production-planning problem includes, among other things, the division of society's scarce inputs among enterprises. Which farm or factory will get how much of which materials? Such decisions can be crucial. If a factory runs short of an essential input, the entire production process may grind to a halt.

As a matter of fact, inputs and outputs cannot be selected separately. The inputs assigned to the production of cars rather than to washing machines determine the quantities of cars and washing machines that can be obtained. However, it is simpler to think of these decisions as if they occur one at a time.

Once again, under laissez-faire it is the price system that apportions fuels and other raw materials among different industries in accord with those industries' requirements. The firm that needs a piece of equipment most urgently will be the last to drop out of the market for that product when prices rise. If more grain is demanded by millers than is currently available, the price will rise and bring quantity demanded back into line with quantity supplied, always giving priority to those users who are willing to pay the most for grain. Thus:

In a free market, inputs are assigned to the firms that can make the most productive (most profitable) use of them. Firms that cannot make a sufficiently productive use of some input will be priced out of the market for that item.

This task, which sounds so simple, is actually almost unimaginably complex. It is also one on which many centrally planned systems have foundered. We will return to it shortly, as an illustration of how difficult it is to replace the market by a central planning bureau. But first let us consider the third of our three coordination problems.

DISTRIBUTION OF PRODUCTS AMONG CONSUMERS

The third task of any economy is to decide which consumer gets each of the goods that has been produced. The objective is to distribute the available supplies so as to match the differing preferences of consumers as well as possible. Coffee lovers must not be flooded with tea while tea drinkers are showered with coffee.

The price mechanism solves this problem by assigning the highest prices to the goods in greatest demand and then letting individual consumers pursue their own self-interests. Consider our example of the rising price of eggs. As the price of eggs rises, those whose craving for omelets is not terribly strong will begin to buy fewer eggs. In effect, the price acts as a rationing device that apportions the available eggs among the consumers who are willing to pay the most for them.

Thus, the price mechanism has an important advantage over other rationing devices: It is able to pay attention to individual consumer preferences. If eggs are rationed by the most obvious and usual means (say, two to a person), everyone ends up with the same quantity—whether he likes eggs or hates them. The price system, on the other hand, permits each consumer to set his own priorities. Thus:

The price system carries out the distribution process by rationing goods on the basis of preferences *and relative incomes*.

Notice the last three words. This rationing process *does* favor the rich, and this is a problem that market economies must confront. However, we may still want to think twice before declaring ourselves opposed to the price system. If equality is our goal, might not a more reasonable solution be to use the tax system to equalize incomes and *then* let the market mechanism distribute goods in accord with preferences?

We have just seen, in broad outline, how a laissez-faire economy addresses the three basic issues of resource allocation: what to produce, how to produce it, and how to distribute the resulting products. Since it performs these tasks quietly, without central direction, and with no apparent concern for the public interest, many radical critics have predicted that such an unplanned system must degenerate into chaos. Yet results from the unplanned market are far from chaotic. Quite ironically, it is the centrally planned economies that have often

found themselves in economic disarray. Perhaps the best way to appreciate the accomplishments of the market is to consider how a centrally planned system must cope with the coordination problems we have just outlined. We will examine just one of them: production planning.

INPUT-OUTPUT ANALYSIS: THE NEAR IMPOSSIBILITY OF PERFECT CENTRAL PLANNING

Of the three coordination tasks of any economy, the assignment of inputs to specific industries and firms has claimed the most attention of central planners. Why? Because the production processes of the various industries are interdependent. Industry X cannot operate without the output of Industry Y, but Y, in turn, finds the product of X indispensable. The whole economy can grind to a halt if the production-planning problem is not solved satisfactorily.

Let's take a simple example. Unless planners allocate enough gasoline to trucking, products will not get to market. Further, unless they allocate enough trucks to haul the gasoline to gas stations, consumers will not be able to get the gasoline. Thus, trucking activity depends on gasoline production but gasoline production also depends on trucking activity. We seem to be caught in a circle. Though it turns out not to be a vicious circle, both truck and gasoline outputs must be decided together, not separately.

Because the output required from any one industry depends on the output desired from every other industry, planners can be sure that the production of the various outputs is sufficient to meet both consumer and industrial demands only by taking explicit account of the interdependencies among industries. If they change the output target for one industry, every other industry's output target also must be adjusted.

For example, if planners decide to provide consumers with more electricity, then more steel must be produced to build more electric generators. But an increase in steel output requires more coal to be mined. More mining, in turn, means that still more electricity is needed to light the mines, to run the elevators, perhaps to operate some of the trains that carry the coal, and so on and on. Any single change in production sets off a chain of adjustments throughout the economy that require still further adjustments.

To decide how much of each output an economy must produce, the planner must use statistics to form a set of equations, one equation for each product, and then solve those equations *simultaneously*. (The simultaneous solution process prevents the circularity of the analysis—electricity output depends on steel production, but steel output depends on electricity production—from becoming a vicious circle.) The technique used to solve these complicated equations—*input-output analysis*—was invented by economist Wassily Leontief, and it won him the Nobel Prize in 1973.

The equations of input-output analysis, which are illustrated in the box on the next page, take account of the interdependence among industries by describing precisely how each industry's target output depends on every other industry's target. Only by solving these equations *simultaneously* for the required outputs of electricity, steel, coal, and so on, can one be sure of a consistent solution that produces the required amounts of each product—including the amount of each product needed to produce every other product.

INPUT-OUTPUT EQUATIONS: AN EXAMPLE

Imagine an economy with only three outputs: electricity, steel, and coal, and let E, S, and C represent the dollar values of their respective outputs. Suppose that to produce every dollar's worth of steel, $0.20 worth of electricity is used up, so that the total electricity demand of steel manufacturers is $0.2S$. Similarly, assume that coal manufacturers use up $0.30 of electricity in producing $1.00 worth of coal, or a total of $0.3C$ units of electricity. Since E dollars of electricity are produced in total, the amount left over for consumers, after subtraction of industrial demands for fuel, will be E (available electricity) minus $0.2S$ (use in steel production) minus $0.3C$ (use in coal production). Suppose further that the central planners have decided to supply $15 million worth of electricity to consumers. We end up with the electricity output equation:

$$E - 0.2S - 0.3C = 15$$

The planner will also need such an equation for each of the two other industries, specifying for each of them the net amount intended to be left for consumers after the industrial uses of the product. The full set of equations might then be:

$$E - 0.2S - 0.3C = 15$$
$$S - 0.1E - 0.06C = 7$$
$$C - 0.15E - 0.4S = 10$$

These are typical equations in an input-output analysis. In practice, however, a typical analysis has dozens and sometimes hundreds of equations with similar numbers of unknowns. This, then, is the logic of input-output analysis.

The example of input-output analysis that appears in the box above is not provided so that you can learn how to apply the technique yourself. Its real purpose is to illustrate the *very complicated* nature of the problem facing a central planner. That problem, while analogous to the one in the box, is enormously more complex. In any real economy, the number of commodities is far greater than the three outputs in the example. In the United States, some large manufacturing companies individually deal in hundreds of thousands of items, and the armed forces keep several *million* different items in inventory.

In planning, it is ultimately necessary to make calculations for each single item. It is not enough to plan the right number of bolts *in total;* we must make sure that the required number *of each size* is produced. (Try to put 5 million large bolts into 5 million small nuts.) To be sure that our plans will really work, we need a separate equation for every size of bolt and one for every size and type of nut. But then, to replicate the analysis described in the boxed insert, we will have to solve simultaneously several *million* equations! This is a task that will strain the capability of the most powerful computer, if it can do the job at all.

Worse still is the data problem. Each of our three equations requires *three* pieces of statistical information, making 3 times 3, or 9 numbers in total. This is because the equation for electricity must indicate on the basis of statistical information how much electricity is needed in steel production, how much in coal production, and how much is demanded by consumers. Therefore, in a five-industry analysis, 5 times 5, or 25 pieces of data are needed; a 100-industry analysis requires 100^2, or 10,000 numbers, and a million-item input-output study might need 1 *trillion* pieces of information. Solving data-gathering problems is, therefore, no easy task, to put it mildly. There are still other complications, but we have seen enough to conclude that:

A full, rigorous central-planning solution to the production problem is a tremendous task, requiring an overwhelming quantity of information and some incredibly difficult calculations. Yet this very complex job is carried out automatically and unobtrusively by the price mechanism in a free-market economy.



In this cartoon from a Soviet humor magazine, one construction worker comments to another, "A slight mistake in the plans, perhaps." It is interesting that there were many cartoons making fun of the inefficiencies of the Soviet economy in the humor magazines of the USSR before the collapse of communism.

HOW PERFECT COMPETITION ACHIEVES EFFICIENCY: WHAT TO PRODUCE

We have now indicated how the market mechanism solves the three basic coordination problems of any economy—what to produce, how to produce, and how to distribute the goods to consumers. Also, we have suggested that these same tasks pose almost insurmountable difficulties for central planners. One critical question remains. Is the allocation of resources that the market mechanism selects *efficient*, according to the precise definition of efficiency presented at the start of this chapter?

The answer, under the idealized circumstances of perfect competition, is yes. Since a detailed proof of this assertion for all three coordination tasks would be long and time-consuming, we will present the proof only for the first of the three tasks—output selection. We will show that, at least in theory, perfect competition does guarantee efficiency in determining the relative quantities of the different commodities that are produced.

We will do this in two steps. First, we will derive a criterion for efficient output selection, that is, a test that tells us whether or not production is being carried out efficiently. Second, we will show that this test is *automatically* passed by the prices that emerge from the market mechanism under perfect competition.

STEP 1: RULE FOR EFFICIENT OUTPUT SELECTION

We begin by stating the rule for efficient output selection:

Efficiency in the choice of output quantities requires that, for each of the economy's outputs, the marginal cost (MC) of the last unit produced be equal to the marginal utility (MU) of the last unit consumed.[2] In symbols:

$$MC = MU$$

[2]Recall from Chapter 5 that we measure marginal utility in money terms, that is, the amount of money that a consumer is willing to give up for an additional unit of the commodity. Economists usually call this the *marginal rate of substitution* between the commodity and money.

This rule is yet another example of the basic principle of marginal analysis that we learned in Chapter 8 (pages 185–201). The goal of the decision on output quantities is to maximize the total benefit (total utility) to society, minus the cost to society of producing the output quantities that are chosen. In other words, we want to maximize the surplus gained by society—total utility minus total cost. But, as was shown in Chapter 8, in order to maximize the difference between total utility and total cost, we must find the outputs that make the corresponding marginal figures—marginal utility and marginal cost—equal to one another. That is what the preceding efficiency rule tells us.

Let us use an example to see explicitly why this rule *must* be satisfied for the allocation of resources to be efficient. Suppose that the marginal utility of an additional pound of beef to consumers is $8, while its marginal cost is only $5. Then the value of the resources that would have to be used up to produce 1 more pound of beef (its MC) would be $3 less than the money value of that additional pound to consumers (its MU). By expanding the output of beef by 1 pound, society could get more (the MU) out of the economic production process than it was putting in (the MC). It follows that the output at which MU exceeds MC cannot be optimal, since society would be made better off by an increase in that output level. The opposite is true if the MC of beef exceeds the MU of beef.

We have therefore shown that, if there is *any* product for which MU is not equal to MC, the economy must be wasting an opportunity to achieve a net improvement in consumers' welfare. This is exactly what we mean by using resources *inefficiently*. Just as was true at point G in Figure 10-1, if MC does not equal MU for any commodity, it is possible to rearrange things so as to make some people better off while harming no one. It follows that efficiency in the choice of outputs is achieved only when MC equals MU for *every* good.[3]

STEP 2: THE CRITICAL ROLE OF THE PRICE SYSTEM

The next step in the argument is to show that under perfect competition, the price system *automatically* leads buyers and sellers to behave in a way that makes MU and MC equal.

To see this, recall from the last chapter that under perfect competition it is most profitable for each beef-producing firm to produce the quantity at which the marginal cost is equal to the price of beef:

$$MC = P$$

This must be so because, if the marginal cost of beef were less than the price, the farmer could add to profits by increasing the size of the herd (or the amount of grain fed to the animals); the reverse would be true if the marginal cost of beef were greater than its price. Thus, under perfect competition, the lure of profits leads each producer of beef (and of every other product) to supply the quantity that makes $MC = P$.

[3]WARNING: As shown in Chapter 13, markets sometimes perform imperfectly because the marginal cost to the decision maker is not the same as the marginal cost to society. This occurs when the individual who creates the cost gets someone else to bear the burden. Example: Firm X's production causes pollution emissions that increase the laundry bills of nearby households. In such a case, Firm X will ignore the cost and produce inefficiently large outputs and emissions.

We also learned, in Chapter 5, that it is in the interest of each consumer to purchase the quantity of beef at which the marginal utility of beef in terms of money is equal to the price of beef:

$$\text{MU} = P$$

If she did not do this, we saw that either an increase or a decrease in her purchase of beef would leave her better off.

Putting these last two equations together, we see that the invisible hand enforces the following string of equalities:

$$\text{MC} = P = \text{MU}$$

But if both the MC of beef and the MU of beef are equal to the same price, *P*, then they must surely be equal to each other. That is, it must be true that the quantity of beef produced and consumed in a perfectly competitive market satisfies the equation:

$$\text{MC} = \text{MU}$$

This is precisely our rule for efficient output selection. Since the same must be true of every other product supplied by a competitive industry:

Under perfect competition, the uncoordinated decisions of producers and consumers can be expected *automatically* and amazingly to produce exactly the quantity of each good that satisfies the MC = MU rule for efficiency. That is, under the idealized conditions of perfect competition, the market mechanism, *without any government intervention* and without anyone else directing it or planning for it, is capable of allocating society's scarce resources efficiently.

THE INVISIBLE HAND AT WORK

This is truly a remarkable result. How can the price mechanism automatically satisfy all of the exacting requirements for efficiency (the marginal utility equals marginal cost condition for each and every commodity)—requirements that no central planner can hope to handle because of the masses of statistics and the enormous calculations they entail? The conclusion seems analogous to the rabbit suddenly pulled from the magician's hat.

But, as always, rabbits come out of hats only if they were hidden there in the first place. What really is the mechanism by which our act of magic works? The secret is that the price system lets consumers and producers pursue their own best interests—something they are probably very good at doing. Prices are the dollar costs of commodities to consumers, so in pursuing their own best interests, consumers will buy the commodities that give them the most satisfaction *per dollar*. As we learned in Chapter 5, this means that each consumer will continue to buy beef until the marginal utility of beef is equal to the market price. Since every consumer pays the same price in a perfectly competitive market, the market mechanism ensures that *every consumer's MU will be equal to this common price.*

Turning next to producers, we know from Chapter 9 that competition equates prices with marginal costs. Once again, since every producer faces the same market price, the forces of competition will bring the MC *of every producer into equality with this common price.* Since MC measures the resource cost (in every firm)

of producing 1 more unit of the good and MU measures the money value (to every consumer) of consuming 1 more unit, then when MC = MU, *the cost of the good to society is exactly equal to the value that consumers place on it*. Thus:

When all prices are set equal to marginal costs, the price system is giving the correct cost signals to consumers. It has set prices at levels that induce consumers to use the resources of society with the same care they devote to watching their own money, because the money cost of a good to the consumer has been set equal to the opportunity cost of the good to society.

This is the magic of the invisible hand. Unlike central planners, consumers need not know how difficult it is to manufacture a certain product or how scarce are the inputs required by the production process. Everything the consumer needs to know to make his or her decision is embodied in the market price, which, under perfect competition, accurately reflects marginal costs.

OTHER ROLES OF PRICES: INCOME DISTRIBUTION AND FAIRNESS

So far we have stressed the role of prices most emphasized by economists: Prices guide the allocation of resources. But a different role of prices often commands the spotlight in public discussions: Prices influence the distribution of income between buyers and sellers. For example, high rents often make tenants poorer and landlords richer.

This rather obvious role of prices draws the most attention from the public, politicians, and regulators, and we should not lose sight of it.[4] Markets serve only those demands that are backed up by consumers' desire *and ability* to pay. Though the market system may do well in serving a poor family, giving that family more food and clothing than a less efficient economy would provide, it offers far more to the family of a millionaire. Many observers object that such an arrangement represents a great injustice, however efficient it may be.

Often, recommendations made by economists for improving the economy's efficiency are opposed on the grounds that they are unfair. For example, economists frequently advocate higher prices for transportation facilities at the time of day when they are most crowded. They propose a pricing arrangement called *peak, off-peak pricing* under which prices for public transportation are higher during rush hours than during other hours.

The rationale for this proposal should be clear from our discussion of efficiency. A seat on a train is a much scarcer resource during rush hours than during other times of the day when the trains run fairly empty. Thus, according to the principles of efficiency outlined in this chapter, seats should be more expensive during rush hours to discourage consumers with a choice from using the trains during peak periods. The same notion applies to other services. Charges for long-distance telephone calls made at night are lower than those in the daytime; in some places, electricity is cheaper at night, when demand does not strain the supplier's generating capacity.

Yet the proposal that higher fares should be charged for public transportation during peak hours—say, from 8:00 A.M. to 9:30 A.M. and from 4:30 P.M. to 6:00 P.M.—often runs into stiff opposition on the grounds that most of the burden will fall on lower-income working people who have no choice about the timing of

[4]Income distribution is the subject of Part IV.

TABLE 10-1 REPLIES TO A QUESTIONNAIRE

QUESTION: In order to make the most efficient use of a city's resources, how should subway and bus fares vary during the day?	Economists (percent)	Conservative Party MPs (percent)	Labour Party MPs (percent)
a. They should be relatively low during rush hour to transport as many people as possible at lower costs.	1	—	40
b. They should be the same at all times to avoid making travelers alter their schedules because of price differences.	4	60	39
c. They should be relatively high during rush hour to minimize the amount of equipment needed to transport the daily travelers.	88	35	19
d. Impossible to answer on the data and alternatives given.	7	5	2

SOURCE: Adapted from Samuel Brittan, *Is There an Economic Consensus?*, p. 93. Copyright Samuel Brittan, 1973. Reproduced by permission of Curtis Brown Ltd.

their trips. For example, a survey in Great Britain of economists and members of Parliament found that while high peak-period fares were favored by 88 percent of the economists, only 35 percent of the Conservative Party MPs and just 19 percent of the Labour Party MPs approved of this arrangement (see Table 10-1 above). We may surmise that the MPs reflected the views of the public more accurately than did the economists. In this case, people simply find the efficient solution unfair, and so they refuse to adopt it.

SAN FRANCISCO BRIDGE PRICING REVISITED

Our earlier example of the bridges in the San Francisco Bay area also raises issues of fairness. Recall that we concluded from our analysis that efficient use of bridges requires higher tolls on the more crowded bridges. Since this principle seems so clear and rational, it may be interesting to see what the actual bridge tolls were in 1995. Travel on the privately owned Golden Gate Bridge required a $3.00 toll for a round trip. But the publicly owned San Francisco–Oakland Bay, Dumbarton, Richmond–San Rafael, and San Mateo–Hayward bridges all carried a uniform $1.00 toll even though, as we saw earlier, average daily traffic on these bridges varies widely, with the San Francisco–Oakland Bay Bridge by far the most crowded.[6]

From the point of view of efficiency, the pattern of tolls on the four publicly owned bridges seems quite irrational. The least crowded bridges were assigned the same tolls as the most crowded bridge! The explanation lies in some widely held notions of fairness.

Many people feel that it is fair for those who travel on a bridge to pay for its costs. In this view, it would be unjust for those who use the crowded San Francisco–Oakland Bay Bridge to pay more, in order to subsidize the least-crowded Dumbarton Bridge. Naturally, a heavily traveled bridge earns more quickly the revenue necessary to recoup the cost of building, maintaining, and running it.

[6]Toll schedule for the San Francisco Bay area bridges provided by Earl Sherman, Caltrans Public Information, California Department of Transportation, December 7, 1995.

On the other hand, the relatively few users of a less crowded bridge must make a fair contribution toward its costs.

Indeed, prices were even further out of line with economists' recommendations at an earlier date. When we first wrote this book, the relatively crowded Golden Gate Bridge cost $1.25, the San Francisco–Oakland Bay, Dumbarton, and San Mateo–Hayward bridges each carried a 75-cent fee, and the sparsely used Richmond–San Rafael Bridge charged a $1 toll.

Of course, such a pattern of tolls does nothing to ease congestion on the overcrowded bridge, and thereby contributes to inefficiency. But one cannot legitimately conclude that advocates of such prices are "stupid." Whether this pattern of tolls is or is not desirable must be decided, ultimately, on the basis of the public's sense of what constitutes fairness and justice in pricing. It also depends on the amount that people are willing to pay in terms of delays, inconvenience, and other inefficiencies in order to avoid apparent injustices.

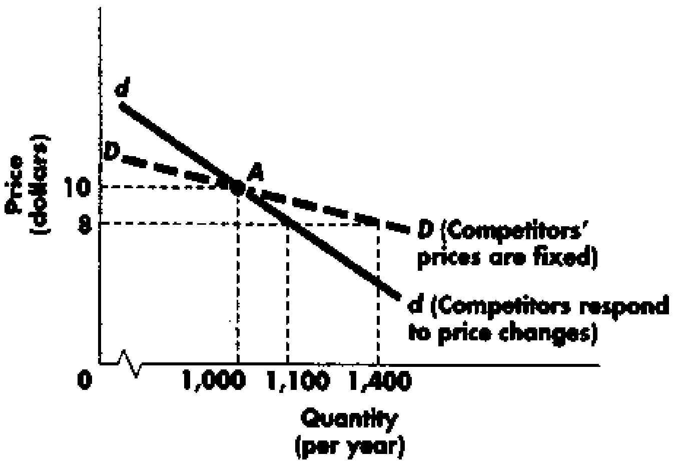
Economics alone cannot decide the appropriate trade-off between fairness and efficiency. It cannot even pretend to judge which pricing arrangements are fair and which are unfair. But it can and should indicate whether a particular pricing decision, proposed because it is considered fair, will impose heavy inefficiency costs upon the community. Economic analysis also can and should indicate how to evaluate these costs, so that the issues can be decided on the basis of an understanding of the facts.

TOWARD ASSESSMENT OF THE PRICE MECHANISM

Our analysis of the case for laissez-faire is not meant to imply that the free-enterprise system is an ideal of perfection, without flaw or room for improvement. In fact, it has a number of serious shortcomings that we will explore in subsequent chapters. But recognition of these imperfections should not conceal the enormous accomplishments of the price mechanism.

We have shown that, given the proper circumstances, it is capable of meeting the most exacting requirements of allocative efficiency, requirements that go well beyond the capacity of any central planning bureau. Even centrally planned economies used the price mechanism to

FIGURE 12-4 THE KINKED DEMAND CURVE



It has been suggested that oligopolists are deterred from changing prices frequently because they fear the reactions of their rivals. If they raise prices, they will lose many customers to competitors because the competitors will not match the price increase. (The elastic demand curve *DD* therefore applies to price increases.) But if they cut prices, competitors will be forced to match the price cut, so the price cut will not bring many new customers. (The inelastic demand curve *dd* applies to price cuts.) Thus, the demand curve facing the firm is the kinked, blue curve *DAd*.

The *DD* curve is the more elastic (flatter) of the two, and a moment's thought indicates why this should be so. If our firm cuts its price from its initial level of $10 to, say, $8, and if competitors do not match this cut, we would expect our firm to get a large number of new customers—perhaps its quantity demanded will jump to 1,400. However, if its competitors respond by also reducing their prices, its quantity demanded will rise by less—perhaps only to 1,100. Conversely, when it raises its price, our firm may expect a larger loss of sales if its rivals fail to match its increase, as indicated by the relative flatness (elasticity) of the curve *DD* in Figure 12-4.

How does this relate to sticky oligopolistic prices? Here we must consider our firm's fears and expectations. The hypothesis of those who designed this model was that a typical oligopolistic firm has good reason to fear the worst. If it lowers its prices and its rivals do not, its sales will seriously cut into its competitor's volume, and so the rivals will *have* to match the price cut in order to protect themselves. The inelastic demand curve, *dd*, will therefore apply if our firm decides on a price reduction (points below and to the right of point *A*).

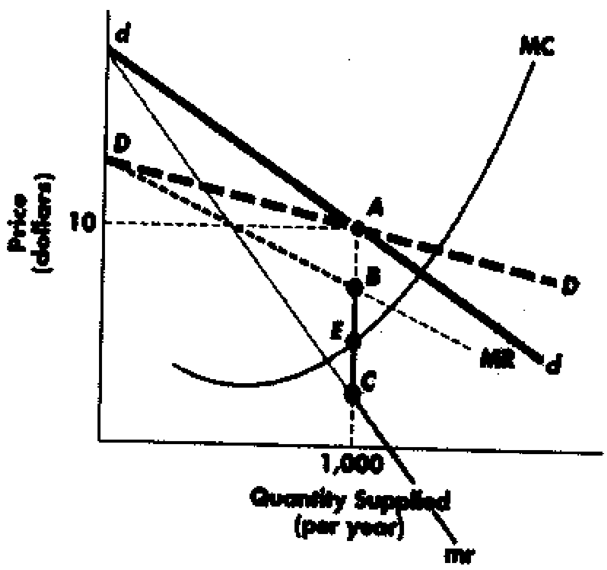
On the other hand, if our company chooses to *increase* its price, management will fear that its rivals will continue to sit at their old price levels, calmly collecting the customers that have been driven to them. Thus, the relevant demand curve for price increases (above *A*) will be *DD*.

In sum, our firm will figure that it will face a segment of the elastic demand curve *DD* if it raises its price and a segment of the inelastic demand curve *dd* if it decreases its price. Its true demand curve will then be given by the heavy blue line, *DAd*. For obvious reasons, this is called a *kinked demand curve*.

The kinked demand curve represents a "heads you lose, tails you lose" proposition in terms of any potential price change. If it raises its price, the firm will lose many customers (because demand is elastic); if it lowers its price, the increase in volume will be comparatively small (because demand is inelastic). In these circumstances, it will pay management to vary its price only under extreme provocation, that is, only if there is an enormous change in costs.

Figure 12-5 illustrates this conclusion graphically. The two demand curves, *dd* and *DD*, are carried over precisely from the previous diagram. The dashed line, labeled MR, is the marginal revenue curve associated with *DD*, while the solid line, labeled mr, is the marginal revenue curve associated with *dd*. The marginal

FIGURE 12-5 THE KINKED DEMAND CURVE AND STICKY PRICES



The kinked demand curve *DAd* that we derived in the previous diagram leads to a marginal revenue curve that follows MR down to point *B*, then drops directly down to point *C*, and finally follows mr thereafter. Consequently, marginal cost curves a little higher or a little lower than the MC curve shown in the diagram will lead to the same price-output decision. Oligopoly prices are "sticky," then, in the sense that they do not respond to minor changes in costs. Only cost changes large enough to push the MC curve out of the range *BC* will lead to changes in price.

revenue curve relevant to the firm's decision making is MR for any output level *below* 1,000 units but mr for any output level *above* 1,000 units. Therefore, the composite marginal revenue curve facing the firm is shown by the thin, blue line *DBC*mr with two angles.

The marginal cost curve drawn in the diagram cuts this composite marginal revenue curve at point *E*, which indicates the profit-maximizing combination of output and price for this oligopolist. Specifically, the quantity supplied at point *E* is 1,000 units, and the price is $10, which we read from demand curve *DAd*.

The unique aspect of this diagram is that the kinked demand curve leads to a marginal revenue curve that takes a sharp plunge between points *B* and *C*. Consequently, moderate upward or downward shifts of the MC curve will still leave it intersecting the marginal revenue curve somewhere between *B* and *C*, and thus will *not* lead the firm to change its output decision. Therefore, *the firm's price will remain unchanged*. (Try this for yourself in Figure 12-5.) This is the sense in which the kinked demand curve makes prices "sticky."

If this is, in fact, the way oligopolists feel about their competitors' behavior, it is easy to see why they may be reluctant to make frequent price changes. We can also understand why a system of price leadership might arise. The price leader can raise prices when he or she thinks it appropriate, confident that the firm will not be left out on a limb (a kink?) by other firms' unwillingness to follow.

THE GAME-THEORY APPROACH

Game theory, contributed in 1944 by mathematician John von Neumann (1903–1957) and economist Oskar Morgenstern (1902–1977), is now the most widely used approach for the analysis of oligopoly behavior. Game theory attacks the issue of interdependence directly by assuming, for the most part, that each firm's managers proceed on the assumption *that their rivals are extremely ingenious decision makers*. In this model, each oligopolist is seen as a competing player in a game of strategy.

Two fundamental concepts of game theory are the *strategy* and the *payoff matrix*. A strategy represents an operational plan for one of the participants. In its simplest form, it may refer to just one possible decision, for example, "I will add a new car to my product line that features a TV set for the driver to watch," or "I will cut the price of my car to $9,500." Since much of the game-theoretic analysis of oligopoly focuses on an oligopoly of two firms—a *duopoly*—we illustrate the payoff matrix for a two-person game in Table 12-1.

This matrix is a table of numbers reporting the profits that each of two rival firms, Firm A and Firm B, can expect to earn—depending on the pricing strategy that each adopts (not knowing the price that the other will offer to customers). The choice open to each firm is either to charge a "high price" or a "low price." The payoff matrix reports the profits that each of the firms can expect to earn, given its own pricing choice and that of its rival. Table 12-1 is read like a mileage chart. For example, the upper, left-hand cell indicates that, if both firms decide to charge high prices, both A and B will earn $10 million.

We also see that, if either firm succeeds in charging a low price when the other does not, the price cutter will actually raise its profit to $12 million (presumably by capturing enough of the market) and drive its rival to a $2 million loss. However, if *both* firms offer low prices, each will be left with a modest $3 million profit.

How does game theory analyze optimal strategic choice? We may envision the management of Firm A reasoning as follows: "If I choose a high-price strategy, the worst that can happen to me is that my competitor will select the low-price counterstrategy, which will cut my return to −$2 million (the orange-shaded number in the first row of the payoff matrix). Similarly, if I select a low-price strategy, the worst possible outcome for me is a $3 million profit (which is the orange-shaded minimum payoff in the second row of the matrix)."

The **MAXIMIN CRITERION** means selecting the strategy that yields the maximum payoff on the assumption that your opponent does as much damage to you as he or she can.

THE MAXIMIN STRATEGY AND THE PRISONERS' DILEMMA

How can the management of Firm A best protect itself from trouble in these circumstances? Game theory suggests that it may be rational to select a strategy based on its *minimum* payoff, just as described above. It should pick the strategy whose *minimum* (that is, its worst) payoff is higher than that for any other strategy: the strategy that offers the highest of the orange-shaded numbers in the matrix. This is called the **maximin criterion**: one seeks the *max*imum of the

TABLE 12-1

A PAYOFF MATRIX

		Firm B's Strategy			
		High Price		Low Price	
Firm A's Strategy	High Price	A gets 10	B gets 10	A gets −2	B gets 12
	Low Price	A gets 12	B gets −2	A gets 3	B gets 3

*min*imum payoffs to the various available strategies. In this case, the maximin strategy for each firm is to offer a low price and earn a profit of $3 million.

Notice that fear of what its rival will do virtually forces each firm to offer a low price and to forgo the high ($10 million) profit that each could earn if it could trust the other to stick to a high price. This example illustrates why many observers conclude that, particularly where the number of firms is small, firms should not be permitted to confer or exchange information on prices. The same sort of analysis also helps to explain how competition limits profits and benefits consumers, and why price cartel arrangements are fragile.

A payoff matrix with a pattern like that in Table 12-1 has many other interesting applications. It is used to show how people can get trapped into making each other (and themselves) worse off, for example by driving polluting cars in the absence of laws requiring emission controls. Each does so because she does not trust other drivers to install emission controls voluntarily. (*EXERCISE:* Make up a payoff matrix that tells this story.)

There is still another interpretation, one which gave this matrix the name by which it is known to game theorists: "the prisoners' dilemma." Here, instead of a two-firm industry, the underlying scenario is that of two burglary suspects who are captured by the police and interrogated in separate rooms. Each suspect has two strategy options: to deny the charge or to confess. If both deny it, both go free, for the police have no other evidence. But if one confesses and the other does not, the silent prisoner can expect the key to his cell to be thrown away. The maximin solution, then, is for both to confess and receive the moderate sentence that this elicits.

OTHER STRATEGIES: THE NASH EQUILIBRIUM

We can interpret the maximin strategy as the pessimist's way of dealing with uncertainty. A player who adopts this strategy assumes that the worst will always happen: no matter what move she makes, her opponent will adopt the countermove that does her the most damage. This type of thinking neglects the possibility that the opponent will not have enough information to find out the most damaging countermove. It also ignores the possibility of finding common ground, as when two competitors collude to extract monopoly profit from consumers.

A **NASH EQUILIBRIUM** results when each player adopts the strategy that gives her the highest possible payoff if her rival sticks to the strategy he has chosen.

There are other strategies that are less pessimistic and still rational. One of the most analytically useful strategies leads to what is called a **Nash equilibrium.** Mathematician John Nash devised this strategy, for which he was awarded the Nobel Prize in economics in 1994. The basic idea is simple. In a two player game, suppose that each is deciding whether to adopt a red or a blue package for its competing product. Assume that each firm earns a higher profit if it selects a package color that is different from the other's. Then, if Firm X happens to select a blue package, it will obviously be most profitable for Y to select a red package. Moreover, it will pay each firm to stick with that choice because red is Y's most profitable response to X's choice of blue, and vice versa.

In general, a Nash equilibrium is a situation in which both players adopt moves such that each player's move is its most profitable response to the move of the other. There are often cases where no such mutually accommodating solution is possible. However, where it is possible, if both players realize this and act accordingly, they may both be able to benefit. Thus, note how much worse off both firms would be in the last example if Firm Y were determined to damage X, whatever the cost to itself, and adopted a blue package, just like X's.

REPEATED GAMES

The scenarios described so far involved one-time transactions, as when a tourist passes through a city and makes a purchase at a store that he will never visit again. Most business transactions are different. A firm usually sells its products day after day, often to repeat buyers. It must continuously review its pricing decisions, knowing that its rival is likely to respond to any change.

A **REPEATED GAME** is one that is played over again a number of times.

Repeated games give players the opportunity to learn something about each other's behavior patterns and, perhaps, to arrive at mutually beneficial arrangements. By adopting some fairly clear pattern of behavior, each firm can seek to attain a reputation that elicits a desired response from its competitor.

We return to the example of the price war between Firm A and Firm B to show how this works. When we studied the payoff matrix for that game, we saw that in a single play in which neither player knew anything about the other's behavior pattern, each player was likely to feel forced to adopt the maximin strategy. In other words, each set a low price for fear that if it adopted the potentially more profitable high price, its rival would adopt a low price and take its customers away. In that way, both firms would end up with low prices and low profits.

When games are repeated, the players may be able to escape such a trap. For example, Firm A can cultivate a reputation for playing a strategy called *tit for tat*. Each time Firm B charges a high price, A responds by charging a high price next time. A follows a similar strategy if B's price is low. After a few repetitions, B will learn that its decisions are always matched by its rival. B will then see that it is better off sticking to a high price. A also benefits from its tit-for-tat approach, which will also lead it, eventually, to permanently high prices.

THREATS AND CREDIBILITY

A player can also use *threats* to induce changes in its rival's behavior. The trouble is that, if carried out, the threat may well damage both parties. For example, a retailer can threaten to double its output and drive prices down near zero if a rival imitates its product. However, the rival is unlikely to believe the threat, since such a low price harms the threatener as much as the threatened. Such a threat is simply not *credible*, with one exception.

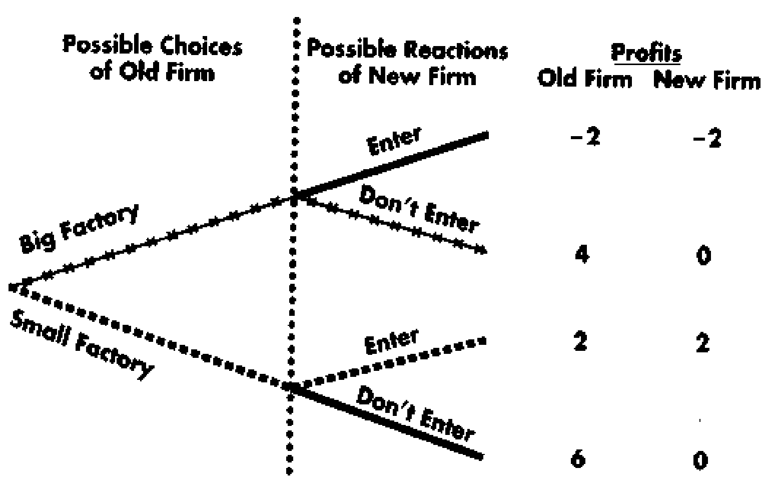
A **CREDIBLE THREAT** is a threat that does not harm the threatener if it is carried out.

The possibility can become a **credible threat** if the threatener takes steps that commit it to carry out the action. For example, if Firm A signed an irrevocable contract that committed it to double its output if B copied A's product, then the threat would become credible, and B would be forced to believe it. There are other ways for A to make a commitment that makes its threat credible. For example, it can build a large plant with plenty of excess capacity. The cost of the factory is high but, once built, that cost is sunk and irrevocable. If the additional cost of turning out the product, once the plant has been built, is very low, then it will pay A to expand its output of the product even at a very low price (if that price exceeds the marginal cost of the item).

This last possibility leads directly to an important application of game theory: the strategic decisions of firms already inside an industry ("old firms") whose primary purpose is to prevent the entry of new firms. Here the old firm considers building a bigger factory than it would otherwise want as a credible threat to potential entrants.

Some hypothetical numbers and a graph typical of those used in game theory will make the story clear. There are two options for the old firm: to build a small factory or a big one. There are also two options for the potential new firm:

FIGURE 12-6 ENTRY AND ENTRY-BLOCKING STRATEGY



This graph shows the possible choices of an existing business (old firm) and the possible responses of a potential entrant (new firm). If the old firm builds a big factory, the entrant will avoid $2 million in losses by staying out of the business, leaving the old firm with $4 million in profit (asterisk lines). On the other hand, with a small factory, the new firm will enter the business (dashed lines) so the old firm will be worse off, with only $2 million in profit.

to open for business (that is, to enter) or not to enter. Figure 12-6 shows the four resulting combinations of decisions that are possible and the corresponding profits or losses that the two firms may expect in each case.

The graph shows that the best outcome for the old firm is if it builds a small factory and the new firm decides not to enter. In that case, the old firm will earn $6 million while the new firm (since it never starts up) will earn zero.

However, if the old firm *does* decide to build a small factory, it can be pretty sure the new firm *will* open up for business, because then the new firm will earn $2 million (rather than zero). In the process, it will reduce the old firm's profit to $2 million.

On the other hand, if the old firm builds a big factory, the increased output will depress prices and profits. The old firm will now earn only $4 million if the new firm stays out, while *each* firm will *lose* $2 million if the new firm enters. Obviously, if the old firm builds a big factory, the new firm will be better off staying out of the business rather than subjecting itself to a $2 million loss.

What size factory, then, should the old firm build? When we consider the firms' interactions, it becomes clear that the old firm should build the large factory with its excessive capacity—for then it can expect the new firm to stay out, leaving the old firm a $4 million profit. Moral: Wasting money on excess capacity may not be wasteful in terms of the oligopolist's self-interest.

There is, of course, a great deal more to game theory than we have explained here. Game theory also provides, for example, an analysis of coalitions. It indicates, for cases involving more than two firms, which firms would do well to align themselves together against which others. The theory of games has also been used to analyze a variety of complicated problems outside the realm of oligopoly theory. It has been employed in management training programs and by a number of government agencies. (See the box, opposite.) It is used in political science and in formulating military strategy.

MONOPOLISTIC COMPETITION, OLIGOPOLY, AND PUBLIC WELFARE

How good or bad, from the viewpoint of the general welfare, is the performance of firms that are monopolistically competitive or oligopolistic?

BILLION-DOLLAR APPLICATION: GAME THEORY AND THE FCC AUCTION



When the Federal Communications Commission announced plans to auction off wavelengths of the radio spectrum for such products as mobile telephones and pagers, many large companies scrambled to hire game theorists to advise them about strategy in this high-stakes undertaking. The FCC used game theory to design the online auction of these so-called *rights to the airways*, and it was up to the bidding companies to figure out how much to pay for the right to service a particular region.

By any measure, the game theorists had their work cut out for them. The bidding was done through each bidder's computer, so checking out the number and size distribution of the competition was difficult. Also, because the commodity being purchased was the latest generation of wireless technology, it was nearly impossible to understand the exact nature of the buying decision. Knowing competitors' cost and revenue structures was another hard one, as was gauging future demand for the service. This enormous uncertainty made nimble strategies all-important.

The FCC might simply have priced the licenses for the various available regions itself. But by conducting an auction, it placed the decision-making onus on the bidding companies and their hired game theorists. Each bidder had to decide what sectors could be most efficiently served, and each had to try and anticipate the most likely moves and countermoves of competitors. After the 3-month-long auction, the FCC had raised nearly $8 billion through the sale of the licenses.

SOURCES: "Winning the Game of Business," *Fortune*, February 6, 1995, p. 36; "Auction Fever," *The Economist*, December 3, 1994, p. 79; William Barnett, "Making Game Theory Work in Practice," *The Wall Street Journal*, February 13, 1995, p. A14; and Edmund L. Andrews, "Winners of Wireless Auction to Pay $7 Billion," *The New York Times*, March 14, 1995, pp. D1 and D3.

We have seen that their performance *can* leave much to be desired. For example, the excess capacity theorem showed us that monopolistic competition can lead to inefficiently high production costs. Similarly, because market forces may not be sufficiently powerful to restrain their behavior, oligopolists' prices and outputs may differ substantially from those that are socially optimal, particularly where the oligopoly organizes itself into a successful cartel. Moreover, there are those who believe that misleading advertising by corporate giants often distorts the judgments of consumers, leading them to buy things they do not need and would otherwise not want. It is said that such corporate giants wield political power, economic power, and power over the minds of consumers—and that all of these undermine the beneficent workings of Adam Smith's invisible hand.

But because oligopoly behavior is so varied, the implications for social welfare differ from case to case. Recent analysis has, however, provided one theoretical case in which both the behavior and the quality of performance of an oligopolistic or monopolistically competitive firm can be predicted and judged unambiguously.[6] It can also serve as a model for government agencies whose job it is to prevent behavior by oligopolistic firms that undermines competition and harms the public interest. This is the case in which entry into or exit from the market is costless and unimpeded. In such a case, called a **perfectly contestable market**, the constant threat of entry forces even the largest firm to behave well—to produce efficiently and never to overcharge. For if that firm is inefficient or sets its prices too high, it will be threatened with replacement by an entrant who offers to serve customers more cheaply.

A market is **PERFECTLY CONTESTABLE** if entry and exit are costless and unimpeded.

A market is defined as perfectly contestable if firms can enter it and, if they choose, exit without losing the money they invested. Note that the crucial issue

[6]See W. J. Baumol, J. C. Panzar, and R. D. Willig, *Contestable Markets and the Theory of Industry Structure*, rev. ed. (San Diego: Harcourt Brace Jovanovich, 1988).

is not the amount of capital that is required to enter the industry, but whether or not an entrant can withdraw the investment if he or she wishes—whether that expenditure is a *sunk* cost. For example, if entry involves investing in highly mobile capital—such as barges, airplanes, or trucks—the entrant may be able to exit quickly and cheaply.[7] If a barge operation decides to serve the lower Mississippi and finds business disappointing, it can easily transfer its boats to, say, the Ohio River.

A profitable market that is contestable is therefore attractive to *potential* entrants. Because of the absence of barriers to entry or exit, firms undertake little risk by going into such a market. If their entry turns out to have been a mistake, they can move to another market without loss.

Because perfect competition requires a large number of firms, all of them negligible in size relative to the size of the industry, no industry with economies of large-scale production can be perfectly competitive. However, markets that contain a few relatively large firms may be highly contestable, though they are certainly not perfectly competitive. But no industry is *perfectly* contestable.

The constant threat of entry elicits good performance by oligopolists, or even by monopolists, in a highly contestable market. In particular, perfectly contestable markets have at least two desirable characteristics.

First, profits exceeding the opportunity cost of capital are eliminated in the long run by freedom of entry, just as they are in a perfectly competitive market. If the current opportunity cost of capital is 12 percent while the firms in a contestable market are earning a return of 18 percent, new firms will enter the market, expand the industry's outputs, and drive down the prices of its products to the point where all excess profit has been removed. To avoid this outcome, established firms must expand output to a level that precludes excess profit.

Second, inefficient enterprises cannot survive in a perfectly contestable industry because cost inefficiencies invite replacement of the incumbents by entrants who can provide the same outputs at lower cost and lower prices. Only firms operating at the lowest possible cost, using the most efficient techniques, can survive. In sum, firms in a perfectly contestable market will be forced to operate as efficiently as possible, and to charge prices as low as long-run financial survival permits.

These ideas have been widely used by courts and government agencies concerned with the performance of business firms (for a recent example, see Chapter 18, page 440). They provide workable guidelines for improved or acceptable behavior in industries in which economies of scale mean that only a small number of firms can or should operate.

A GLANCE BACKWARD: COMPARISON OF THE FOUR MARKET FORMS

We have now completed the set of chapters that has taken us through the four main market forms that characterize the economy: perfect competition, monopoly, monopolistic competition, and oligopoly. You have probably absorbed a lot of information about the workings of these market forms as you read Chapters 9 through 12, but you may be confused by the details. Table 12-2 presents an

---

[7] Earlier, it was widely thought that air transportation is a highly contestable industry, but recent evidence suggests that while this judgment is not entirely incorrect, it requires considerable reservations.

TABLE 12-2 ATTRIBUTES OF THE FOUR MARKET FORMS

Market Form	Number of Firms in the Market	Frequency in Reality	Entry Barriers	Public Interest Results	Long-Run Profit	Equilibrium Conditions
Perfect competition	Very many	Rare (if any)	None	Good	Zero	MC = MR = AC = AR = P
Pure monopoly	One	Rare	Likely to be high	Misallocates resources	May be high	MR = MC
Monopolistic competition	Many	Widespread	Minor	Inefficient	Zero	MR = MC AR = AC
Oligopoly	Few	Produces large share of GDP	Varies	Varies	Varies	Vary

overview of the main attributes of each of the market forms for comparison. It shows:

1. Perfect competition and pure monopoly are concepts useful primarily for analytical purposes. Neither of these is found very often in reality. Monopolistically competitive firms occur in profusion, and oligopoly firms account for the largest share of the economy's output.
2. Profits are zero in long-run equilibrium under perfect competition and monopolistic competition because entry is so easy, so that high profits attract new rivals into the market.
3. Consequently, AC = AR in long-run equilibrium under these two market forms. In equilibrium, MC = MR for the profit-maximizing firm under any market form. However, under oligopoly, firms may adopt the strategies described by game theory or they may pursue goals other than profits; for example, they may be sales maximizers. Therefore, in the equilibrium of the oligopoly firm, MC may be unequal to MR.
4. The point is that the behavior of the perfectly competitive firm and industry theoretically leads to an efficient allocation of resources that maximizes the benefits to consumers, given the resources available to consumers. The same is, incidentally, theoretically true for any market form if the market is perfectly contestable. But otherwise, monopoly misallocates resources to a greater or smaller degree by restricting its output in order to raise prices and profits. Under monopolistic competition, excess capacity and inefficiency are apt to result. Under oligopoly, almost anything can happen, so it is impossible to generalize about its vices or virtues.

SUMMARY

1. Under **monopolistic competition**, there are numerous small buyers and sellers; each firm's product is at least somewhat different from every other firm's product—that is, each firm has a partial "monopoly" of some product characteristics, and thus a downward-sloping demand curve; there is freedom of entry and exit; and there is perfect information.
2. In long-run equilibrium under monopolistic competition, free entry eliminates economic profits by forcing the firm's demand curve into a position of tangency with its average cost curve. Therefore, output will be below the point at which average cost is lowest. This is why monopolistic competitors are said to have "excess capacity."

3. An oligopolistic industry is composed of a few large firms selling similar products in the same market.
4. Under **oligopoly,** each firm carefully watches the major decisions of its rivals and often plans counterstrategies. As a result, rivalry is often vigorous and direct, and the outcome is difficult to predict.
5. One model of oligopoly behavior assumes that the oligopolists ignore interdependence and simply maximize profits or sales. Another assumes that they join together to form a **cartel** and thus act like a monopoly. A third possibility is **price leadership,** where one firm sets prices and the others follow suit.
6. A firm that maximizes sales will continue producing up to the point where marginal revenue is driven down to zero. Consequently, a sales maximizer will produce more than a profit maximizer and will charge a lower price.
7. If a firm thinks that its rivals will match any price cut but fail to match any price increase, its demand curve becomes "kinked" and its price will be sticky—that is, it will be adjusted less frequently than would be the case under either perfect competition or pure monopoly.
8. *Game theory* provides new tools for the analysis of business strategies under conditions of oligopoly.
9. In a **maximin strategy,** the player takes the strongest possible precautions against the worst possible outcome of any move it selects.
10. A **Nash equilibrium** is one in which each player adopts the move that yields the highest possible payoff to itself, given the move selected by the other player.
11. In **repeated games,** a firm can seek to acquire a reputation that induces the other player to make decisions that do not damage its interests. It may also promote its goals by means of **credible threats.**
12. Monopolistic competition and oligopoly can be harmful to the general welfare. But since behavior varies widely, the implications for social welfare also vary from case to case.

KEY TERMS

Monopolistic competition
Excess capacity theorem
Oligopoly
Oligopolistic interdependence
Cartel
Price leadership
Sales maximization
Kinked demand curve
Sticky price
Game theory
Price war
Maximin criterion
Nash equilibrium
Prisoners' dilemma
Repeated game
Credible threat
Perfectly contestable market

QUESTIONS FOR REVIEW

1. How many real industries can you name that are oligopolies? How many that operate under monopolistic competition? Perfect competition? Which of these is hardest to find in reality? Why do you think this is so?
2. Consider some of the products that are widely advertised on TV. By what kind of firm is each produced—a perfectly competitive firm, an oligopolistic firm, or what? How many major products can you think of that are *not* advertised on TV?
3. In what ways may the small retail sellers of the following products differentiate their goods from those of their rivals to make themselves monopolistic competitors: hamburgers, radios, cosmetics?
4. Pricing of securities on the stock market is said to be done under conditions in many respects similar to perfect competition. The auto industry is an oligopoly. How often do you think the price of a share of Ford Motor Company's common stock changes? How about the price of a Ford Taurus? How would you explain the difference?
5. Suppose that Chrysler hires a popular singer to advertise its compact automobiles. The campaign is very successful, and the company increases its share of the compact-car market substantially. What is Ford likely to do?
6. Using game theory, set up a payoff matrix similar to one Chrysler's management might employ in analyzing the problem presented in Review Question 5.
7. Review Question 4 at the end of Chapter 11 presented cost and demand data for a monopolist and asked you to find the profit-maximizing solution. Use these same data to find the sales-maximizing solution. Are the answers different? Explain.
8. A new entrant, Bargain Airways, cuts air fares between Eastwich and Westwich by 20 percent. Biggie Airlines, which has been operating on this route,

responds by cutting fares by 35 percent. What does Biggie hope to achieve?

9. If air transportation were perfectly contestable, why would Biggie fail to achieve the ultimate goal of its price cut?
10. Which of the following industries are most likely to be contestable?
    a. Aluminum production
    b. Barge transportation
    c. Automobile manufacturing

    Explain your answers.
11. Since the deregulation of air transportation, a community served by a single airline is no longer protected by a regulatory agency from monopoly pricing. What market forces, if any, restrict the ability of the airline from raising prices as a pure monopolist would? How effective do you think those market forces are in keeping air fares down?
12. Explain, for a repeated game (a) why it may be advantageous to have the reputation of a tough guy who always takes revenge against anyone who harms your interests? (b) Why it may be advantageous to have a reputation of irrationality?

CHAPTER 13

THE MARKET MECHANISM: SHORTCOMINGS AND REMEDIES

When she was good
She was very, very good,
But when she was bad
She was horrid.
Henry Wadsworth Longfellow

What does the market do well, and what does it do poorly? This issue is the focus of our microeconomic analysis, and we are well on our way toward answers. Chapters 9 and 10 explained the workings of Adam Smith's invisible hand—the mechanism by which a perfectly competitive economy allocates resources efficiently without any guidance from government. While that model is just a theoretical ideal, observation of reality confirms the accomplishments of the market mechanism. Free-market economies have achieved levels of output, productive efficiency, variety in available consumer goods, and general prosperity that are unprecedented in history—and are now the envy of the formerly planned economies.

Yet the market mechanism has its weaknesses. In Chapters 11 and 12, we examined one of these—its vulnerability to exploitation by large and powerful business firms, which leads both to concentration of wealth and to misallocation of resources. Now we take a more comprehensive view of the failures of the market and of some of the things that can be done to remedy them. Clearly, the market cannot do everything we want. Amid the outpouring of products, we find areas of appalling poverty, cities choked by traffic and pollution, and educational institutions and artistic organizations in serious financial trouble. Though our economy produces an overwhelming abundance of material wealth, it seems far less able to reduce social ills and environmental damage. We will examine the reasons for these failings and indicate why the price system *by itself* may not be able to deal with them.

Failure to recognize the shortcomings of the market mechanism and the fact that it cannot produce instantaneous miracles has already been a source of disillusionment in the countries of eastern Europe. Many had unrealistic expectations of immediate economic prosperity, with little or no social costs, as soon as the market mechanism was introduced.

However, recognition of the limitations of the market does not imply that the public interest calls for its abandonment. As we will see, many of the imperfections of this economic system seem to be amenable to treatment within the market environment, sometimes even making use of the market mechanism to cure its own deficiencies.

PUZZLE: RISING HEALTH-CARE COSTS IN CANADA

Long before the U.S. government made a serious attempt to grapple with the costs of health care, the Canadians adopted a universal health-care program intended to solve the problem. For this purpose, strong controls over prices and fees were imposed. Each Canadian province has one insurance plan that reimburses doctors according to a uniform fee schedule; hospitals are put on a predetermined overall budget; patients pay very low direct, out-of-pocket costs.

There are many observers who believe the Canadians have created an efficient, user-friendly system, though it does have some critics. But what is clear is that the Canadians have not succeeded in containing costs. Despite price controls, Canadian health-care costs have been rising persistently faster than the rate of inflation, just as in the United States, where there are no such national rules to rein in rising prices for health-care services. Does this mean that Canadian services are especially inefficient or corrupt? There is no evidence for such suspicions. Then why should the Canadians have failed to brake their cost of health care? The materials in this chapter will help you to understand the answer, with its significant implications for U.S. policy.

WHAT DOES THE MARKET DO POORLY?

While an exhaustive list of its imperfections is not possible, we can list some major areas in which the market has been accused of failing:

1. Market economies suffer from severe business fluctuations.
2. The market distributes income rather unequally.
3. Where markets are monopolized, they allocate resources inefficiently.
4. The market deals poorly with the side effects of many economic activities.
5. The market cannot readily provide public goods, such as national defense.
6. The market may do a poor job of allocating resources between the present and the future.
7. The market mechanism makes public and personal services increasingly expensive, and this often induces socially damaging countermeasures by government.

The first three items in the list are studied elsewhere in this book. This chapter deals with the remaining four. To help us analyze these cases, we briefly review the concept of efficient resource allocation, discussed in detail in Chapter 10.

EFFICIENT RESOURCE ALLOCATION: A REVIEW

The basic problem of resource allocation is deciding how much of each commodity should be produced by the economy. At first glance, it may seem that

the solution is simple: the more, the better; we should produce as much of each good as we can. But careful thought tells us that this is not necessarily so.

Outputs are not created out of thin air. They are produced from scarce supplies of labor, fuel, raw materials, and machinery. If we use these resources to produce, say, more jeans, we must take them away from some other products, such as backpacks. To decide whether increasing the production of jeans is a good idea, we must compare the utility of that increase with the loss of utility caused by having to produce fewer backpacks. It is *efficient* to increase the output of jeans only if society considers the additional jeans more valuable than the forgone backpacks.

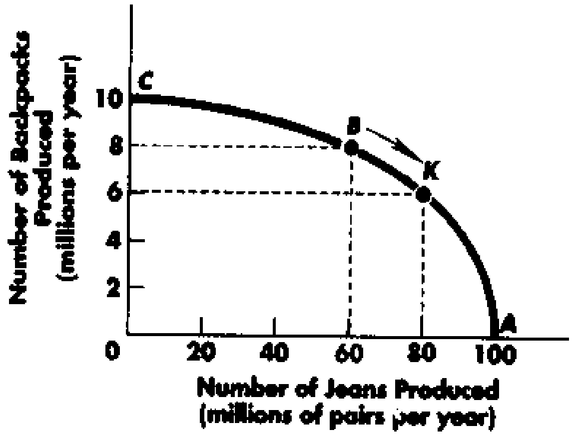
OPPORTUNITY COST AND RESOURCE ALLOCATION

Here we recall the concept of *opportunity cost*, one of our **Ideas for Beyond the Final Exam.** The opportunity cost of an increase in the output of some product is the value of the other goods and services that must be forgone when inputs (resources) are taken away from their production in order to increase the output of the product in question. In our example, the opportunity cost of the increased jeans output is the decrease in output of backpacks that results when resources are reallocated from the latter to the former. The general principle is that an increase in some output represents a *misallocation* of resources if the utility of that increased output is less than its opportunity cost.

To illustrate this idea, we repeat a graph encountered several times in earlier chapters—a *production possibilities frontier*—but we put it to a somewhat different use. Curve *ABC* in Figure 13-1 is a production possibilities frontier showing the alternative combinations of jeans and backpacks that the economy can produce by reallocating its resources between the production of the two goods. Suppose that point *B*, representing the production of 8 million backpacks and 60 million pairs of jeans, represents the *optimal* resource allocation. We thus assume it is the only combination of outputs that best satisfies the wants of society among all the possibilities that are *attainable* (given the technology and resources as represented by the production frontier). Two questions are pertinent to our discussion of the price system:

FIGURE 13-1 THE ECONOMY'S PRODUCTION POSSIBILITIES FRONTIER FOR THE PRODUCTION OF TWO GOODS

This graph shows the combinations of outputs of the two goods that the economy can produce. Let *B* be the most desired output combination. If the price of a backpack is above its marginal cost, or the price of a pair of jeans is below its marginal cost, then backpack output will be inefficiently small and jeans output inefficiently large (point *K*).



1. What prices will get the economy to select point *B;* that is, what prices will yield an *efficient* allocation of resources?
2. How can the wrong set of prices lead to a misallocation of resources?

The first question was discussed extensively in Chapter 10, where we saw that:

An efficient allocation of resources requires that each product's price be equal to its marginal cost; that is:

$$P = MC$$

The reasoning, in brief, is as follows. In a free market, the price of any good reflects the money value to consumers of an additional unit, that is, its *marginal utility* (MU). Similarly, if the market mechanism is working well, the *marginal cost* (MC) measures the value (the opportunity cost) of the resources needed to produce an additional unit of the good. Hence, if prices are set equal to marginal costs, then consumers, by using *their own money* in the most effective way to maximize *their own* satisfaction, will automatically be using *society's resources* in the most effective way. In other words, as long as it sets prices equal to marginal costs, the market mechanism automatically satisfies the MC = MU rule for efficient resource allocation that we studied in Chapter 10.[1] In terms of Figure 13-1, this means that if $P = MC$ for both goods, the economy will automatically gravitate to point *B*, which we assumed to be the optimal point.

This chapter is devoted mainly to the second question: How can the "wrong" prices cause a *mis*allocation of resources? The answer to this question is not too difficult, and we can use the case of monopoly as an illustration.

The "law" of demand tells us that a rise in the price of a commodity normally will reduce the quantity demanded. Suppose, now, that the backpack industry is a monopoly, so the price of backpacks exceeds their marginal cost.[2] This will decrease the quantity of backpacks demanded below the 8 million that we have assumed to be socially optimal (point *B* in Figure 13-1). The economy will move from point *B* to a point such as *K*, where too few backpacks and too many pairs of jeans are being produced for maximal consumer satisfaction. By setting the "wrong" prices, then, the market prevents achievement of the most efficient use of the economy's resources.

If the price of a commodity is above its marginal cost, the economy will tend to produce less of that item than the amount necessary to maximize consumer benefits. The opposite will occur if an item's price is below its marginal cost.

In the remainder of this chapter, we will encounter several other instances in which the market mechanism may set the "wrong" prices.

EXTERNALITIES: GETTING THE PRICES WRONG

We come now to the fourth item on our list of market failures, the first studied in this chapter. It is one of the least obvious, yet one of the most consequential

[1] If you need review, consult pages 245–248.

[2] To review why price under monopoly may be expected to exceed marginal cost, you may want to reread pages 263–265.

of the imperfections of the price system. Many economic activities provide incidental benefits to others for whom they are not specifically intended. For example, homeowners who plant beautiful gardens in front of their homes incidentally and unintentionally provide pleasure to their neighbors and to those who pass by—people from whom they receive no payment. We say then that their activity generates a **beneficial externality.**

An activity is said to generate a **BENEFICIAL or DETRIMENTAL EXTERNALITY** if that activity causes incidental benefits or damages to others, and no corresponding compensation is provided to or paid by those who generate the externality.

Similarly, there are activities that indiscriminately impose costs on others. For example, the operators of a motorcycle repair shop, from which all sorts of noise besieges the neighborhood and for which they pay no compensation to others, are said to produce a **detrimental externality.** Pollution constitutes the classic illustration of a detrimental externality.

To see why the presence of externalities causes the price system to misallocate resources, we need only recall that the system achieves efficiency by rewarding producers who serve consumers well—that is, at as low a cost as possible. This argument breaks down, however, as soon as some of the costs and benefits of economic activities are left out of the profit calculation.

When a firm pollutes a river, it uses up some of society's resources just as surely as when it burns coal. However, if the firm pays for coal but not for the use of clean water, it is to be expected that management will be economical in its use of coal and wasteful in its use of water. Similarly, a firm that provides benefits to others for which it receives no payment is unlikely to be generous in allocating resources to the activity, no matter how socially desirable it may be.

In an important sense, the source of the market mechanism's difficulty here lies in society's rules about property rights. Coal mines are *private property;* their owners will not let anyone take coal without paying for it. Thus, coal is costly and so is not used wastefully. But waterways are not private property. Since they belong to everyone in general, they belong to no one in particular. They therefore can be used free of charge as dumping grounds for wastes by anyone who chooses to do so. Because no one pays for the use of the dissolved oxygen in a public waterway, that oxygen will be used wastefully. That is the source of detrimental externalities—the fact that waterways are exempted from the market's normal control procedures.

EXTERNALITIES AND INEFFICIENCY

Using these concepts, we can see precisely why an externality has undesirable effects on the allocation of resources. In discussing externalities, it is crucial to distinguish between *social* and *private* marginal cost. We define **marginal social cost (MSC)** as the sum of two components: (1) **marginal private cost (MPC),** which is the share of marginal cost caused by an activity that is paid for by the persons who carry out the activity, and (2) *incidental cost,* which is the share borne by others.

The **MARGINAL SOCIAL COST (MSC)** of an activity is the sum of its **MARGINAL PRIVATE COST (MPC)** plus the incidental cost (positive or negative) which is borne by others.

If increased output by a firm increases the smoke that the firm emits, then, in addition to its direct private costs as recorded in the company accounts, expansion of production imposes incidental costs on others. These costs take the form of increased laundry bills, medical expenditures, outlays for air conditioning and electricity, as well as the unpleasantness of living in a cloud of noxious fumes. These are all part of the activity's marginal *social* cost.

Where the firm's activities generate detrimental externalities, its marginal social cost will be greater than its marginal private cost. In symbols, MSC > MPC. Therefore, the firm's output must be too big. This must be so because, in equilibrium, the market will yield an output at which consumers' marginal

utility (MU) is equal to the firm's marginal private cost (MU = MPC). It follows that the marginal utility is *smaller* than marginal social cost. Society would then necessarily benefit if output of that product were *reduced*. It would lose the marginal utility but save the marginal social cost. We conclude that:

Where the firm's activity causes detrimental externalities, free markets will leave us in a situation where marginal benefits are less than marginal social costs. Smaller outputs than those that maximize profits will be socially desirable.

We have already indicated why this is so. Private enterprise has no motivation to take into account costs that it causes to others but for which it does not have to pay. Goods that cause such externalities will be produced in undesirably large amounts by private firms. For precisely analogous reasons:

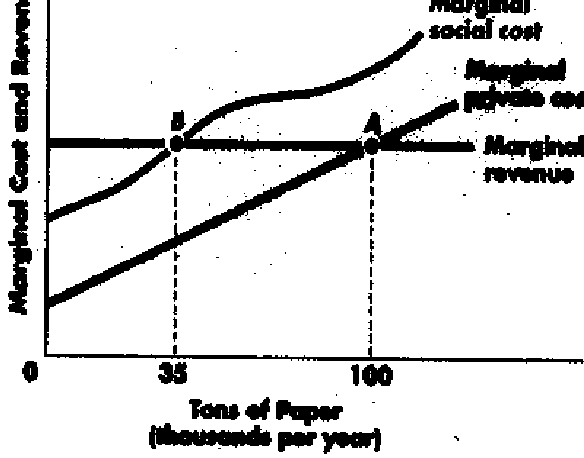
Where the firm's activity generates beneficial externalities, free markets will produce too little output. Society would be better off with larger output levels.

These principles can be illustrated with the aid of Figure 13-2. This diagram repeats the two basic curves needed for the analysis of the equilibrium of the firm: a marginal revenue curve and a marginal cost curve (see Chapter 8). These curves represent the *private* costs and revenues accruing to a particular firm (in this case, a paper mill). The mill maximizes profits by providing 100,000 tons of output corresponding to the intersection between the marginal cost and marginal revenue curves (point *A*).

Now suppose that the factory's wastes pollute a nearby waterway, so that its production creates a detrimental externality whose cost the owners do not pay. Then marginal social cost must be higher than marginal private cost, as shown in the diagram. The output of paper, which is governed by private costs, will be 100,000 tons (point *A*)—an excessive amount from the viewpoint of the public interest, given its environmental consequences.

Notice that if, instead of being able to impose the external costs on others, the paper mill's owners were forced to pay them, then their private marginal cost curve would correspond to the higher of the two cost curves. Paper output

FIGURE 13-2 EQUILIBRIUM OF A FIRM WHOSE OUTPUT PRODUCES A DETRIMENTAL EXTERNALITY (POLLUTION)



The firm's profit-maximizing output, at which its marginal private cost and its marginal private revenue are equal, is 100,000 tons. But if the firm paid all the social costs of its output instead of shifting some of them to others, its marginal cost curve would be the curve labeled "Marginal social cost." Then it would pay the firm to reduce its output to 35,000 tons, thereby reducing the pollution it caused.

would then fall to 35,000 tons, corresponding to point *B*, the intersection between the marginal revenue curve and the marginal *social* cost curve.

The same sort of diagram can be used to show that the opposite relationship will hold when the firm's activity produces beneficial externalities. The firm will produce less of its beneficial output than it would if it were rewarded fully for the benefits that its activities yield. Beneficial externalities arise when the activities of Firm A create incidental benefits for Firm B or Individual C (and perhaps for many others as well), or when A's activities *reduce* the costs of others' activities. For example, Firm A's research laboratories, while making its own products better, may also incidentally discover new research techniques that reduce the research costs of other firms in the economy.

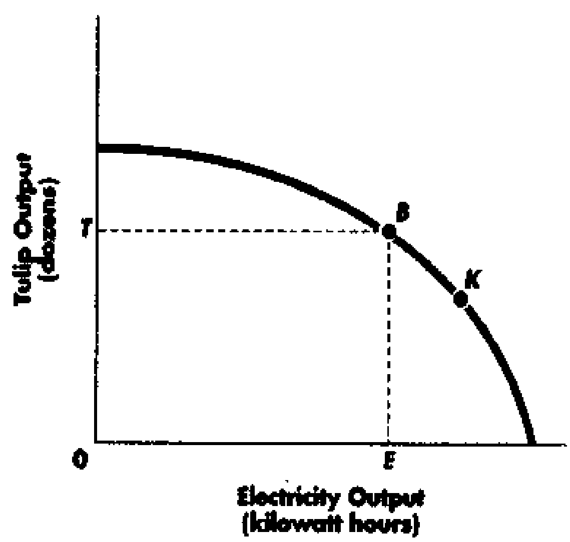
But these results can perhaps be seen more clearly with the help of a production possibilities frontier diagram similar to that in Figure 13-1. In Figure 13-3, we see the frontier for two industries: electricity generation, which causes air pollution (a detrimental externality), and tulip growing, which makes an area more attractive (a beneficial externality). We have just seen that detrimental externalities make marginal social cost greater than marginal private cost. Hence, if the electric company charges a price equal to its own marginal (private) cost, that price will be less than the true marginal social cost. Similarly, in tulip growing, a price equal to marginal private cost will be above the true marginal cost to society.

We saw earlier in the chapter that an industry that charges a price above marginal cost will reduce quantity demanded through this high price, and so it will produce an output too small for an efficient allocation of resources. The opposite will be true for an industry whose price is below marginal social cost. In terms of Figure 13-3, suppose that point *B* again represents the efficient allocation of resources, involving the production of *E* kilowatt hours of electricity and *T* dozen tulips.

Because the polluting electric company charges a price below marginal social cost, it will sell more than *E* kilowatt hours of electricity. Similarly, because tulip growers generate external benefits, and so charge a price above marginal social

FIGURE 13-3 EXTERNALITIES, MARKET EQUILIBRIUM, AND EFFICIENT RESOURCE ALLOCATION

Because electricity producers emit smoke (a detrimental externality), they do not bear the true marginal social cost of their output. Therefore, electricity price will be below marginal social cost, and electricity output will be inefficiently large (point K, not point B). The opposite is true of tulip production. Because they generate beneficial externalities, tulips will be priced above their marginal social cost and tulip output will be inefficiently small.



cost, they will produce less than T dozen tulips. The economy will end up with the resource allocation represented by point K rather than that at point B. There will be too much smoky electricity production and too little attractive tulip growing. More generally:

An industry that generates detrimental externalities will have a marginal social cost higher than its marginal private cost. If its price is equal to its own marginal private cost, it will therefore be below the

CHAPTER 14

REAL FIRMS AND THEIR FINANCING: STOCKS AND BONDS

The action of the stock market must necessarily be puzzling at times since otherwise everyone who studies it only a little bit would be able to make money in it.
B. Graham, D. L. Dodd, and S. Cottle

Earlier chapters have provided a theoretical analysis of the business firm's decisions. But a firm does more than select inputs, outputs, and prices. In this chapter, we look at some other salient features of real firms.

We begin by describing the different types of firms that make up U.S. business. Then we describe the most important ways in which firms acquire resources for investment. This leads us to look at the stock and bond markets, to which many individuals bring money, hoping to make it grow.

PUZZLE: THE STOCK MARKET'S UNPREDICTABILITY

The stock market is something of an enigma. No other economic activity is reported in such detail in so many newspapers and followed with such concern by so many people; yet few activities have so successfully eluded prediction of their future. There is no shortage of well-paid "experts" prepared to forecast the future of the market or a particular stock. But there are real questions about what these experts deliver.

For example, a widely noted study of leading analysts' predictions of company earnings (on which they based their stock-price forecasts) reports:

> . . . [W]e wrote to 19 major Wall Street firms . . . among the most respected names in the investment business.
>
> We requested—and received—past earnings predictions on how these firms felt earnings for specific companies would behave over both a 1-year and a 5-year period. These estimates . . . were . . . compared with actual results to see how well the analysts forecast short-run and long-run earnings changes. . . .
>
> Bluntly stated, the careful estimates of security analysts (based on industry studies, plant visits, etc.) do very little better than those that would be obtained by simple extrapolation of past trends. . . .
>
> For example . . . the analysts' estimates were compared [with] the assumption that every company in the economy would enjoy a growth in earnings approximating the long-run rate of growth of the national income. It often turned out that . . . this naïve forecasting model . . . would make smaller errors in forecasting long-run earnings growth than . . . [did] the professional forecasts of the analysts. . . .
>
> When confronted with the poor record of their 5-year growth estimates, the security analysts honestly, if sheepishly, admitted that 5 years ahead is really too far in

advance to make reliable projections. They protested that, while long-term projections are admittedly important, they really ought to be judged on their ability to project earnings changes 1 year ahead.

Believe it or not, it turned out that their 1-year forecasts were even worse than their 5-year projections.[1]

It has been said that an investor may as well pick stocks by throwing darts at the stock market page—it is far cheaper to buy a set of darts than to obtain the apparently useless advice of a professional analyst. Indeed, there have been at least two experiments, one by a U.S. senator and one by *Forbes* magazine, in which stocks picked by dart throwing actually outperformed the mutual funds, whose stocks are selected by experts.

Later in this chapter we will suggest an explanation of this poor performance.

FIRMS IN THE UNITED STATES

Firms typically are divided into three groups: *corporations, partnerships,* and *individual proprietorships* (businesses having single owners). The importance of corporations in the U.S. economy is suggested by the fact that their sales amount to about two-thirds of the country's GDP. Almost all large American firms are corporations. General Motors by itself sold $155 billion in 1994, and Exxon and Ford each sold over $100 billion. The combined sales of these three firms alone amount to considerably more than the GDPs of Austria, Belgium, the Netherlands, Sweden, Switzerland, and many, many more countries. Most industries in which these giant firms are found are *oligopolies*, a market form we analyzed in Chapter 12.

But while a huge chunk of America's output comes from corporations, less than 20 percent of American business firms are incorporated. The reason is that most firms are small. Even corporations are often quite small—about 40 percent of them have business receipts of less than $100,000. But by far the greatest number of firms (counting all firms large and small, and including the corner grocery store and shoe repair shop) are individual or family proprietorships. Of the 20 million business firms in the United States, about 15 million are proprietorships, 3.7 million are corporations, and 1.5 million are partnerships.

The nation's small business firms have a disproportionately small share of U.S. business. These small firms have earnings that are not only relatively low but also very risky—risky in the sense that the average new firm does not last very long (its average life is reported to be less than 7 years). When making economic decisions, the buyer is not the only one who must beware!

Let us take a closer look at the basic forms of business organization, and what induces organizers of a firm to choose one form rather than another.

PROPRIETORSHIPS

A **PROPRIETORSHIP** is a business firm owned by a single person.

Most small retail firms, farms, and many small factories are **proprietorships.** A proprietorship involves fewer legal complications than any other form of business organization. To start a proprietorship, an individual simply decides to go

[1] Burton G. Malkiel, *A Random Walk Down Wall Street* (New York: W. W. Norton, 1990), pp. 140–141.

into business and to open a new firm or take over an existing one. This is a major advantage of the proprietorship form of organization.

But its main attraction is probably that the owner can be his or her own boss and the firm's sole decision maker. No partners or stockholders have to be consulted when the proprietor wants to expand or change the company's product line or modify the firm's advertising policy. A proprietorship also has tax advantages, particularly compared with a corporation. A proprietor's income is taxed only once. If the same firm were to incorporate, its income would be taxed twice—once as the income of the firm (the corporate income tax) and again as the personal income of the owner.

UNLIMITED LIABILITY is a legal obligation of a firm's owner(s) to pay back company debts with whatever resources he or she owns.

On the other hand, a proprietorship has a disadvantage that makes it almost impossible to organize large-scale enterprises as proprietorships: the owner has **unlimited liability** for the debts of the firm. If the company goes out of business leaving unpaid bills, the former owner can be forced to pay them out of personal savings. The owner can be made to sell the family home, private collections of stamps or paintings, or any other personal assets, no matter how unrelated to the business, so that the proceeds can be used to pay off the company's debts.

SUMMARY

There are three main advantages of the individual proprietorship:

1. It leaves full control in the hands of the owner.
2. It involves little legal complication.
3. It generally reduces the taxes its owners must pay.

Its two main disadvantages are:

1. The unlimited liability of the owner for the debts of the company.
2. The difficulty of raising substantial funds for the firm.

PARTNERSHIPS

A **PARTNERSHIP** is a firm whose ownership is shared by a fixed number of proprietors.

Measured in terms of the amounts of their capital, **partnerships** tend to be larger than proprietorships but smaller than corporations. However, the largest partnerships greatly exceed the smallest corporations in terms of both their financing and their influence. For example, some of the most prestigious law firms and investment banks are partnerships.

The advantage of the partnership over the proprietorship is that it combines the funds and expertise of a number of people to form a company larger than any one of the owners could have financed or managed alone. A partnership may also bring together a variety of specialists, as often happens in a medical practice. In addition, the partnership offers the advantage of freedom from double taxation, a benefit it shares with the proprietorship.

But the partnership has disadvantages, some of them substantial. Decision making in a partnership may be harder than in any other type of firm. The sole proprietor need consult no one before acting; the corporation appoints officers who are authorized to decide things for the company. But in a partnership, it may be necessary for every partner to agree before any steps are taken by the firm, and this is the primary bane of this form of enterprise. A partnership has been compared to two people in a horse costume, each supplying two of the

legs, each prepared to go in a different direction, and each unable to move without the other.

Furthermore, partners, like sole proprietors, have unlimited liability. They can conceivably be in danger of losing their personal possessions to pay off company debts. Finally, the partnership suffers from unique legal complications. A partnership agreement is like a marriage contract entered into solely for the financial advantage of the participants, and so there is likely to be considerable haggling about the terms. Also, under the law, if a partner dies or decides to leave the firm, the partnership may have to be dissolved and haggling about the contract may begin anew.

SUMMARY

The benefits of the partnership to the owners of the firm are:

1. Access to larger quantities of capital.
2. Protection from double taxation.

Its disadvantages are:

1. The need to obtain the agreement of many if not all partners to all major decisions.
2. Unlimited liability of the partners for the obligations of the company.
3. The legal complications, including automatic dissolution of the partnership, when there is *any* change in ownership.

CORPORATIONS

A **CORPORATION** is a firm that has the legal status of a fictional individual. This fictional individual is owned by a number of persons, called its *stockholders*, and is run by a set of elected officers and a board of directors, whose chairman is often also in a powerful position.

LIMITED LIABILITY is a legal obligation of a firm's owners to pay back company debts only with the money they have already invested in the firm.

Most big firms are **corporations,** a form of business organization that has quite a different legal status from that of a proprietorship or a partnership. Though it seems strange, a corporation is an individual in the eyes of the law. Therefore, its earnings, like those of other individuals, are taxed. This leads to double taxation of the stockholders. That is, corporate earnings are taxed twice, once when they are earned by the company, and a second time when they go to the investors in the form of dividends and are subject to the ordinary income tax on the investors' income.

But this disadvantage is counterbalanced by an important advantage: any debt of the corporation is regarded as an obligation of that fictitious individual, not as a liability of any stockholder. This means that the stockholders benefit from the protection of **limited liability**—they can lose no more than the money they have put into the firm.

Limited liability is the main secret of the success of the corporate form of organization. Thanks to that provision, individuals from every part of the world are willing to put money into firms whose operations they do not understand and whose managements they do not know. Each shareholder receives in return a claim on the firm's profits, and, at least in principle, a portion of the company's ownership.

Corporations are directed by hired groups of managers: a chairman of the board of directors, a president, various vice presidents, and so on. These executives are, legally, employees of the owners of the firm who, as we will see, are the stockholders of the corporation. This arrangement has great advantages. It prevents the quarrels and indecision that are often problems for partnerships.

On the other hand, since the management is made up of hired personnel, it cannot always be trusted to do what is best for the owners. Managers are often accused of looking after their own interests first and, if necessary, sacrificing those of the stockholders (the owners). This is a problem that has recently received much attention in discussions of takeovers of corporations—that is, purchase of control of the firm by a group of outsiders—a subject we will examine later in this chapter.

Corporations escape one other problem that troubles partnerships. As we saw, if a partner wants to leave the firm, the entire enterprise may have to be reorganized. But in a corporation, any owner who wants to quit just sells her stock and the corporation goes on exactly as before. In this way, at least in theory, a corporation can continue forever.

SUMMARY

Benefits of the corporate form to the owners:

1. Limited liability.
2. Access to large quantities of capital.
3. Ease of operation with the help of a hired management.
4. "Permanence": The firm is not dissolved or reorganized each time an owner leaves.

Its disadvantages are:

1. Double taxation of payments to the owners.
2. The possibility that hired managers will act in their own interests rather than those of the owners.

EFFECT OF DOUBLE TAXATION OF CORPORATE EARNINGS

Does an investor end up earning less by putting money in a corporation than by putting it in a company that is about equally risky but not subject to double taxation? Paradoxically, the answer is that investors, on the average, will *not* lose anything by choosing the corporation. The tax will not and cannot put those who make one type of investment at a disadvantage in comparison with those who choose any other.

How is this possible? How does the effect of the additional tax on a corporate stock disappear before it reaches the stockholder? There are two processes that achieve this act of magic.

First, corporations are forced to avoid some investment opportunities that partnerships and proprietorships can afford to take on. Suppose that the market rate of return to people who provide money to firms is 9 percent, and a new product is invented that is expected to bring a 12 percent return to a firm that manufactures it. An individual proprietor can afford to produce the new item—borrowing the necessary funds at 9 percent and keeping the 3 percent additional return on the new item for herself. But a large corporation *cannot* afford to produce the new item. In order to compete for funds, it must also pay investors 9 percent, which means that it will have to earn about 15 percent on its investments since about 40 percent of that money will be siphoned off in corporate taxes.

Thus, double taxation keeps corporate business out of various economic activities that offer real, but limited, earnings potential. This effect may be

unfortunate from the viewpoint of the efficiency of the economy, because it means that many firms are induced to stay out of activities in which it might be useful for them to take part. For instance, corporations may find it too costly to open retail outlets in slum areas or to run trains to isolated rural areas—activities that might be profitable in the absence of the tax.

There is a second fail-safe mechanism that protects new investors in corporate stocks from earning lower returns on the average than they would on other securities of equal risk. Suppose that two otherwise identical securities, A and B, each offer a return of $60 per year, but A is subject to a 50 percent tax while B is not. *Question:* If the market price of Security B is $1,000, what will be the market price of A? *Answer:* The price of A will be only $500, exactly half the price of Security B. Why? Because it will bring in only $30 per year after taxes, exactly half of what Security B returns, investors will be willing to pay only half as much for it as they are willing to pay for the untaxed security. But at those prices, investors in either security will obviously earn the same rate of return after payment of taxes.

Double taxation of corporate earnings tends to restrict the activities of corporate firms, keeping them out of relatively low-profit operations. However, double taxation does not mean that the individual investor earns less by putting money into a corporation than by putting it into other businesses.

STOCKS AND BONDS

Our discussion of the earnings of an investor in corporate securities introduces a subject of interest to millions of Americans—*stocks* and *bonds*, the financial instruments that provide funds to the corporate sector of the economy.

A **COMMON STOCK** of a corporation is a piece of paper that gives the holder of the stock a share of the ownership of the company.

A **BOND** is simply an I.O.U. by a corporation that promises to pay the holder of the piece of paper a fixed sum of money at the specified *maturity* date and some other fixed amount of money (the *coupon* or the *interest payment*) every year up to the date of maturity.

Common stock represents ownership of part of a corporation. For example, if a company issues 100,000 shares, then a person who owns 1,000 shares actually owns 1 percent of the company and is entitled to 1 percent of the company's *dividends*, which are the corporation's annual payments to stockholders. The shareholder's vote counts for 1 percent of the total votes in an election of corporate officers or in a referendum on corporate policy.

Bonds differ from stocks in several ways. First, whereas the purchaser of a corporation's stock *buys* a share of its ownership and receives some control over its affairs, the purchaser of a bond simply *lends* money to the firm. Second, whereas stockholders have no idea how much they will receive for their stocks when they sell them, or how much they will receive in dividends each year while they own them, bondholders know with a high degree of certainty how much money they will be paid if they hold their bonds to maturity. For instance, a bond with a face value of $1,000, with an $80 coupon that matures in 2004, will provide to its owner $80 per year every year until 2004, and in addition, it will repay the $1,000 to the bondholder in 2004. Unless the company goes bankrupt, there is no doubt about this repayment schedule. Third, bondholders legally have a *prior claim* on company earnings, which means that nothing can be paid by the company to its stockholders until interest payments to the company's bondholders have been met. For all these reasons, bonds are considered less risky to their buyers than stocks.

An important exception are "junk bonds"—very risky bonds that became popular in the 1980s. They were used heavily by groups of persons trying to purchase enough stocks of some corporation to acquire control of that firm. We

will say more later about such "takeover" activities and the use of junk bonds to finance them.

In reality, the differences between stocks and bonds are not as clear-cut as just described. Two relevant misconceptions are particularly worth noting. First, the ownership of the company represented by the holding of a few shares of its stock may be more symbolic than real. A holder of 0.002 percent of the stock of AT&T—which is a *very large* investment—exercises no real control over AT&T's operations.

In fact, many economists believe that the ownership of large corporations is so diffuse that stockholders or stockholder groups rarely have any effective control over management. In this view, the management of a corporation is a largely independent decision-making body; as long as it keeps enough cash flowing to stockholders to prevent discontent and *organized* rebellion, management can do anything it wants within the law. Looked at in another way, this last conclusion really says that stockholders are merely another class of persons who provide loans to the company. The only real difference between stockholders and bondholders, according to this interpretation, is that stockholders' loans are riskier and therefore entitled to higher payments.

Second, bonds *can* be quite risky to the bondholder. Persons who try to sell their bonds before maturity may find that the market price for bonds happens to be low, so that if they need to raise cash in a hurry, they may incur substantial losses. Also, bondholders may be exposed to losses from inflation. Whether the $1,000 promised the bondholder at the 2004 maturity date represents substantial purchasing power or only a little depends on what happens to the general price level in the meantime. No one can predict the price level this far in advance with any accuracy. Finally, a firm can issue bonds for which there is little backing; that is, the firm may own little valuable property that it can use as a guarantee of repayment to the lender—the bondholder. This has often been true of the junk bonds of the 1980s, and it helps to explain their high riskiness.

BOND PRICES AND INTEREST RATES

Why is investment in bonds risky? What makes their prices go up and down? The main element in the answer is that changes in interest rates cause bond prices to change. There is a straightforward relationship between bond prices and interest rates: whenever one goes up, the other *must* go down.

For example, suppose that Sears, Roebuck had issued some 15-year bonds when interest rates were comparatively low, so that the company had to pay only 6 percent to sell bonds. People who invested $1,000 in new Sears bonds then received in return a contract that promised them $60 per year for 15 years plus the return of their $1,000 at the end of that period. Suppose now that interest rates in the economy rise, so that new 15-year bonds of companies of similar quality pay 12 percent. Now an investor with $1,000 can buy a contract that offers $120 per year. Obviously, no longer will anyone pay $1,000 for a bond that promises only $60 per year. Consequently, the market price of the old Sears bonds must fall.

This example is not hypothetical. Until a few years ago there were bonds in existence that had been issued much earlier at interest rates of 6 percent and even less. In the 1980s' markets, with interest rates well above 6 percent, such bonds sold for prices far below their original values.

When interest rates in the economy rise, there must be a fall in the prices of previously issued bonds with their lower interest earnings. For the same reason,

when interest rates in the economy fall, the prices of previously issued bonds must rise.

It follows that as interest rates in the economy change because of changes in monetary policy or other reasons, bond prices fluctuate. That is one reason why investment in bonds can be risky.

FINANCING CORPORATE ACTIVITY

When a corporation needs money to add to its plant or equipment or to finance other types of real investment, it can get it by printing new stock certificates or new bonds and selling them to people who are looking for something in which to invest their money. What enables the firm to get money in exchange for printed paper? Doesn't the process seem a bit like counterfeiting? If done improperly, there are grounds for the suspicion. But, carried out appropriately, it is a perfectly rational economic process.

As long as the funds derived from a new issue of stocks and bonds are used effectively to increase the firm's capacity to produce and earn a profit, then these funds will automatically yield the means for any required repayment and for the payment of appropriate amounts of interest and dividends to the purchasers. But there are times when this does not happen. It is alleged that one of the favorite practices of the more notorious 19th-century manipulators of the market was "watering" of company stocks—issuing stocks with little or nothing to back them up. The term is derived from the practice of some cattle dealers who would force their animals to drink large quantities of water just before bringing them to be weighed for sale.

PLOWBACK (or RETAINED EARNINGS) is the portion of a corporation's profits that management decides to keep and invest back into the firm's operations rather than to pay out directly to stockholders in the form of dividends.

Another major source of funds is **plowback or retained earnings.** For example, if a company earns $30 million after taxes and decides to pay out only $10 million in dividends and invest the remaining $20 million back into the firm, that $20 million is called "plowback."

When business is profitable and management has the funds to reinvest in the company, it will often prefer plowback to other sources of funding. One reason for this preference is that it is usually less risky to management. This source of funds, unlike others, does not require prior scrutiny by the Securities and Exchange Commission (SEC), the government agency that regulates stocks.[2] Moreover, plowback does not depend on the availability of eager customers for new company stocks and bonds. An issue of new securities can be a disappointment if there is little public demand for them when they are offered. But plowback runs no such risk.

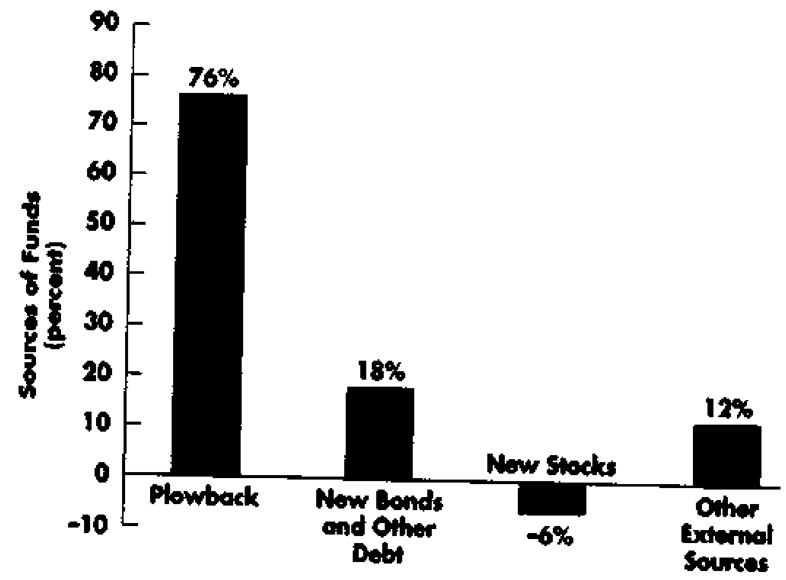
Above all, a plowback decision generally does not call attention to the degree of success of management's operations as a new stock issue does. In these instances, the SEC, potential buyers of the stock, and their professional advisers may all scrutinize the company carefully.

Another reason for the attractiveness of plowback is that issuing new stocks and bonds is usually an expensive and lengthy process. The company is required by the SEC to gather masses of data in its prospectus—a document describing the financial condition of the company—before the new issue is approved.

---

[2]The Securities and Exchange Commission, established in 1934, protects the interests of people who buy securities. It requires firms that issue stock and other securities to provide information about their financial condition, and it regulates the issue and trading of securities.

FIGURE 14-1 SOURCES OF NEW FUNDS, U.S. CORPORATIONS, 1994



Corporations in the United States get about 76 percent of their reinvestment funds from plowback, which consists mostly of depreciation—funds accumulated for replacement of plant, equipment, and so on, as it wears out or becomes obsolete.

SOURCE: U.S. Bureau of the Census, *Statistical Abstract of the United States 1995*, 115th ed. (Washington, D.C.: U.S. Government Printing Office, September 1995).

A final way for a company to obtain money is by borrowing it from banks, insurance companies, or other private firms with money to lend. It may also sometimes borrow from a U.S. government agency, either directly or with the agency's help. (The agency serves as guarantor in this instance, promising to make sure the loan is repaid.) For example, loans may be arranged with the help of the national defense agencies if they want a private firm to undertake the design and production of an expensive new weapons system. Small business firms, too, are eligible for various forms of assistance in borrowing.

Figure 14-1 (a bar chart) shows the relative importance of each of the different sources of funds to U.S. corporations. It indicates that plowback is by far the most important source of corporate financing, constituting about 76 percent of the total financing to the corporate sector of the economy in 1994. Issues of new bonds and other forms of borrowing supplied about 18 percent of the total, and sales of stock actually contributed a *negative* amount of funding (−6 percent), as corporations reduced the number of their stocks in the hands of the public by buying some of them back.

CORPORATE CHOICE BETWEEN STOCKS AND BONDS

We have seen why a corporation may prefer to finance its real investment, such as construction of factories and equipment, through plowback or retained earnings rather than through the issue of new stocks or bonds. But suppose it has decided to do the latter. How does it determine whether bonds or stocks suit its purposes better?

Two considerations are of prime importance. Although issuing bonds generally exposes the firm to more risk than issuing stocks, the corporation usually expects to pay more money to stockholders over the long run than to bondholders. In other words, to the firm that issues them, bonds are cheaper but riskier. The decision about which is better for the firm therefore involves a tradeoff between the two considerations.

Why are bonds risky to the corporation? When it issues $20 million in new bonds at 10 percent, the company commits itself to pay out $2 million every year for the life of the bond, whether the business is booming or losing money. That is a big risk. If the firm is unable to meet its obligation to bondholders in some year, it faces bankruptcy.

The issue of new stocks does not burden the company with any such risk since the company does not promise to pay the stockholders *any* fixed amount. Stockholders simply receive whatever is left of the company's net earnings after payments to bondholders. If nothing is left to pay the new stockholders in some years, legally speaking, that is just their bad luck. The higher risk faced by stockholders is the reason they normally obtain higher average expected payments from the company than bondholders.

To the firm that issues them, bonds are riskier than stocks because they commit the firm to make a fixed annual payment, even in years when it is losing money. For the same reason, stocks are riskier than bonds to the buyers of securities. That is why stockholders expect to be paid more money than bondholders.

BUYING STOCKS AND BONDS

Although stocks and bonds can be purchased through any brokerage firm, not all brokers charge the same fees. Bargain brokerage houses advertise in the financial pages of newspapers, offering investors very little service—no advice, no research, no other frills—other than merely buying or selling what the customer wants them to, at lower fees than those charged by higher-service brokerage firms.

Many investors are not aware of the various ways in which stocks can be purchased (or sold). Two noteworthy arrangements are (a) a *market order* purchase, which simply tells the broker to buy a specified quantity of stock at the best price the market currently offers, and (b) a *limit order*, which is an agreement to buy a given amount of stock when its price falls to a specified level. If the investor offers to buy at $18, then shares will be purchased by the broker if and when the market price falls to $18 per share or less.

A recent survey of shareowners by the New York Stock Exchange (NYSE) estimated that 47 million individuals—perhaps one out of five people in the United States—owned stock in a publicly traded company or in a stock mutual fund. The NYSE estimates that an additional 100 million people also participate in the stock market indirectly through banks, pension funds, insurance companies, and the like. Such organizations, called *institutional investors*, are estimated to own over half of all stocks traded on the NYSE.

SELECTING A PORTFOLIO: DIVERSIFICATION

A set of various securities owned by a person or an organization is called the owner's *portfolio* of investments. A portfolio may well be far less risky than any of the individual securities it contains. The secret is **portfolio diversification**, not putting all of one's eggs in one basket.

If Joe Jones divides his holdings among Companies A, B, and C, then the portfolio may perform satisfactorily, even if Company A goes broke. Moreover, suppose that Company A specializes in producing luxury items, which do well in prosperous periods but very badly during recessions, while Company B sells cheap clothing, whose cyclical demand pattern differs greatly from that of Company A. If Jones holds stock in both companies, the overall risk is less than if he

PORTFOLIO DIVERSIFICATION means including a number and variety of stocks, bonds, and other such items in an individual's portfolio. If the individual owns airline stocks, for example, diversification requires the purchase of a stock or bond in a very different industry, such as a breakfast cereal producer.

owned stock in only one of the companies. All other things being equal, a portfolio containing many different types of securities tends to be less risky than a portfolio with fewer types of securities.

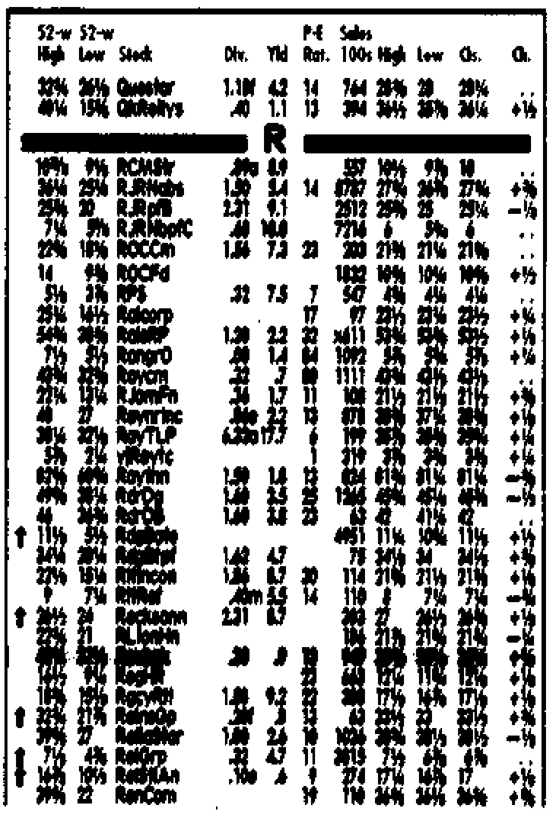
Increasingly, institutional investors have adopted portfolios composed of broad ranges of stocks typifying those offered by the entire stock market. By owning a representative basket of stocks, money managers can reduce the risks of owning individual stocks and ensure that their portfolios are not significantly outperformed by the overall market.

Institutional money managers have been making increasing use of computers to decide on their portfolios and to buy or sell huge portfolios of stocks simultaneously and rapidly. Since 1982, some traders have also allowed their computers to decide when to jump in and make massive sales or purchases. This is called *program trading*. During 1995, program trading accounted for about 11 percent of total NYSE volume and a considerable amount elsewhere. Program trading is controversial, and some observers have argued that it has aggravated price fluctuations, especially during the stock market crash of October 1987.

FOLLOWING A PORTFOLIO'S PERFORMANCE

Newspapers carry daily information on stock and bond prices. Figure 14-2 is an excerpt from the stock market page of *The Denver Post*. In the first two columns, before the company name, the report gives the stock's highest and lowest price in the last year. In the highlighted example, the price of Reebok stock is reported to have ranged between $\$40\frac{3}{4}$ and $\$31\frac{1}{8}$. After the name of the stock, there appears

FIGURE 14-2 EXCERPT FROM A STOCK-PRICE TABLE



This table from *The Denver Post* gives the highest and lowest price in the current year; the name of the stock; the current dividend rate; the dividend yield; the ratio of stock price to company earnings (P/E ratio); the number of shares sold; the highest, lowest, and final price on the previous day; and the change in price from the day before.

SOURCE: *The Denver Post*, August 18, 1995, p. 4C.

FIGURE 14-3 EXCERPT FROM A BOND-PRICE TABLE

This report from *The New York Times* shows the name of the bond; the annual interest payment; the year in which the bond will be redeemed (that is, the year in which the company will repay that debt); the yield (that is, the annual payment per dollar of current market price); and the previous day's closing price of the bond, as well as the change in price from the day before.

SOURCE: *The New York Times*, October 12, 1995, p. D20.

Company	Cur.Yld	Vol	Price	Chg
Fldcst 6s12 cv	...	10	84	+ ½
FstUnRE 8⅞03	8.8	20	100½	...
FrepIM 10⅞01	10.5	30	103¾	...
FremGen	...	35	40¼	+1
GMotAc 8¼16	8.0	16	103	...
GMotAc dc6s11	6.9	36	87	− ¼
GMotAc 8⅝97	8.2	7	102⅝	−1⅜
GMotAc zr12	...	67	287¾	− ¼
GMotAc zr15	...	34	239½	+ ½
GnHost 8s02 cv	8.6	3	92½	−1¼
GnHost 11½02	11.5	75	99½	...
Genrad 7¼11 cv	7.7	47	94	+ ½
GaPow 6⅛99	6.3	5	98	−1½
GPA Del 8¾98	...	12	88¼	− ⅜
Grancre 6½03	...	262	88	+ ½
Guilrd 6s12	6.2	3	96¾	...
Hallwd 7s00	...	74	80	+ ¼
Halwd 13½09	29.3	1	46⅛	−1⅞
ICN Phrm 8½99	...	15	110½	+2½
ICN Phrm 12⅞98	12.5	6	103⅝	...
IMC Glb 6¼01	...	98	110½	+3
IBM 8⅜19	7.3	110	114	...
IBM 6⅜97	6.3	176	100¾	− ¼
IBM 7¼02	6.9	6	104½	...
IBM 6⅜00	6.3	115	100½	− ⅛
IBM 7½13	7.2	22	104⅜	− ⅛
IntTch 9⅜96	9.4	45	100	+ 11/32
IRT Pr 7 3s03	...	35	91	−1

the annual dividend per share ($.30). Following that is the yield, or the dividend as a percent of the closing price (.9 percent).

The next column reports the *price earnings* (P/E) ratio (13 for Reebok). This latter figure is the price per share divided by the company's net earnings per share in the previous year, and it is usually taken as a basic measure indicating whether the current price of the stock overvalues or undervalues the company. However, no simple rule enables us to interpret the P/E figures—for example, a very risky firm or a slowly growing firm with a low P/E may be considered overvalued, while a safe, rapidly growing firm with a high P/E may still be a bargain.

The next column indicates the number of shares that were traded on the previous day (94,900), an indication of whether that stock is actively traded. Finally, the last four figures indicate yesterday's highest price ($$35\frac{3}{8}$), its lowest price ($$35\frac{1}{8}$), the price at which the last transaction of the day took place ($$35\frac{1}{4}$), and the change in that price from the previous day ($+\frac{1}{4}$).

Figure 14-3, this time from *The New York Times*, gives similar information about bonds. The first thing to notice here is that a given company may have several different bonds differing in maturity date and coupon (annual interest payment). For example, IBM offers five different bonds. The one that is highlighted is labeled IBM 7½13, meaning that these bonds pay an annual interest rate of 7.5 percent (the coupon) on their face value and that their maturity (redemption) date is 2013. Next, the current yield is reported as 7.2 percent. This is simply the coupon divided by the price. Since that yield, 7.2 percent, is lower than the coupon, 7.5 percent, the bond must be selling at a price above its face value, so that the return per dollar is correspondingly lower. The remaining information in the table means the same as that reported for stock prices.

STOCK EXCHANGES AND THEIR FUNCTIONS

The *New York Stock Exchange*—the "Big Board"—is the most prestigious stock market. Located at the beginning of Wall Street in New York City, it is "*the* establishment" of the securities industry. Only the best-known and most heavily traded securities are dealt with by the New York Stock Exchange, which handles over 2,000 stocks. The leading brokerage firms hold "seats" on the stock exchange, which enable them to trade directly on the floor of the exchange. Altogether, the exchange has over 600 member organizations. Seats are traded on the open market; in October of 1995, for example, a seat on the New York Stock Exchange went for $1 million.

Someone who wants to buy a stock on the New York Stock Exchange must use a broker who will deal with a firm that has a seat on the exchange. Suppose that you live in Ohio and want to buy 200 shares of General Motors. The broker you approach may be employed by a firm that holds a seat on the exchange, or she may work through another firm that holds one. The broker who is to fill your order contacts a person called a *specialist*, who works on the floor of the exchange and who handles GM stock.

The specialist usually owns some GM stock of his own that he will offer to sell if no other sellers are available at the moment. In addition, the specialist usually receives a number of limit orders from investors offering to sell specified quantities of GM stock at specified prices. There may, for example, be one offer to sell 5,000 shares at any price above $55 and another offer to sell 1,200 shares at any price above $60. Similarly, the specialist is likely to have several limit orders to buy at various specified prices.

The floor broker brings your order to the specialist, who determines a price that, in his judgment, more or less balances supply and demand as indicated by his recent sales and purchases and the limit orders in his possession. At this price, the specialist will fill your order from one of the limit orders to sell (he must do so whenever possible), or he will fill it from his personal inventory of General Motors stock. The price determination process we have just described is sometimes called the *auction market* process.

The New York Stock Exchange expedites this "auction" by using an elaborate electronic system to link member firms directly to the appropriate specialists or floor brokers. This system handles approximately 75 percent of all NYSE orders.

While 85 percent of all stock market transactions (in dollar volume) are handled by the NYSE, the *American Stock Exchange*, located a few blocks away, trades many stocks that are heavily demanded but that are not exchanged in quite as large volumes as those handled by the Big Board. About 2.5 percent of the dollar volume of stock trades occurs on the American Stock Exchange. There are also *regional exchanges*—such as the Midwest, Cincinnati, Pacific Coast, Philadelphia, and Boston exchanges—which deal in many of the same stocks that are handled on the New York Stock Exchange. A good portion of the business of regional exchanges, like that of the New York Stock Exchange, is serving large "institutional" customers such as banks, insurance companies, and mutual funds. Their volume amounts to about 12.5 percent of the total stock traded.

A growing share of total stock trading is carried out "over the counter" via the institution called *NASDAQ* (the automated quotation system of the National Association of Securities Dealers), which carries out its operations electronically. NASDAQ handles the stocks of more than 4,900 companies, including such giants as Intel and Microsoft.

Second, the stock market determines the current price of the company's

YOU ARE THERE

AN EVENT ON THE TRADING FLOOR OF THE NEW YORK STOCK EXCHANGE

The trading floor of the New York Stock Exchange is a crowded set of rooms cluttered with milling people, hundreds of computer monitors, and other paraphernalia. It is a very high-tech space in a 93-year-old architectural relic of a bygone era. The floor contains many stations, or "posts," each of which is presided over by a *specialist*, who is assigned responsibility for trading a particular set of stocks.

The floor's frenetic activity suddenly focuses on one specialist's post. News has just come in that one of the companies whose stock this specialist handles has earned more in the last quarter than has been expected. Brokers crowd around the specialist with orders to buy and sell the company's stock, as its price rises rapidly in the wake of the good news. Deals are completed verbally. The specialist's clerk records the trades and enters them into the computerized tape, where they are instantly available for all to see anywhere in the world.

SOURCE: Murray Teitelbaum, Communications Division, New York Stock Exchange, January 1996.



NASDAQ is a "dealer market" in which a number of securities dealers compete with one another in the sale and purchase of stocks. More than 400 such dealer firms are active "market makers" on NASDAQ, buying and selling for their own accounts. More than 10 dealers may trade an average NASDAQ stock, and more than 40 dealers compete for the most traded stocks. The dealers charge no commissions and post (electronically) the prices at which they are willing to buy and sell a given stock. The sources of the market makers' earnings is the excess of the selling price over the buying price, called *the spread*. They also profit from changes in the prices of the securities they hold.

REGULATION OF THE STOCK MARKET

The U.S. securities markets are regulated by both the government and the industry itself. At the base of the regulatory pyramid, brokerage firms maintain compliance departments to oversee their own operations. At the next level, the New York Stock Exchange, the American Stock Exchange, NASDAQ, and the regional exchanges are responsible for monitoring the business practices, the adequacy of funding, and the compliance and integrity of their member firms. They also utilize sophisticated computer surveillance systems to scrutinize trading activity. The Securities and Exchange Commission (SEC) is the federal government agency that oversees the market's self-regulation.

One example of these self-imposed rules is the steps that the markets have taken since the October 1987 market crash to cushion such price falls. The NYSE

and the Chicago Mercantile Exchange adopted a series of "coordinated circuit breakers," which halt all equities trading for 1 hour if the Dow Jones Industrial Average falls 250 points from the previous day's close. Trading would be halted for an additional 2 hours if the Dow were to fall another 150 points on the same day. However, no one is sure whether these and other similar measures will prove very effective in preventing sharp drops in stock prices.

STOCK EXCHANGES AND CORPORATE CAPITAL NEEDS

While corporations often raise the funds they need by selling stock, they do not normally do so through any of the stock exchanges. New stock issues usually are handled by a special type of bank called an *investment bank*. In contrast, the stock markets trade almost exclusively in "secondhand securities"—stocks in the hands of individuals and others who had bought them earlier and who now wish to sell them.

Thus, the stock market does not provide funds to corporations that need financing to expand their productive activities. The markets provide money only to persons who already hold stocks previously issued by the corporations.

Yet stock exchanges have two functions that are of critical importance for the financing of corporations. First, by providing a secondhand market for stocks, they make it much less risky for an individual to invest in a company. Investors know that their money is not locked in—if they need the money, they can always sell their stocks to other investors or to the specialist at the price that the market currently offers. This reduction in risk makes it far easier for corporations to issue new stocks.

Second, the stock market determines the current price of the company's stocks. That, in turn, determines whether it will be hard or easy for a corporation to raise money by selling new stocks.

Some people believe that the price of a company's stock is closely tied to the efficiency of its operations, its effectiveness in meeting consumer demands, and its diligence in going after profitable innovation. In this view, those firms that can make effective use of funds because of their efficiency are precisely the corporations whose stock prices will usually be comparatively high. In this way the stock market tends to channel the economy's investment funds to those firms that can make best use of the money. In sum:

If a firm has a promising future, its stock will tend to command a high price on the stock exchanges. The high price of its stock will make it easier for the firm to raise capital by permitting it to amass a large amount of money through the sale of a comparatively small number of new shares of stock. Thus, *the stock market helps to allocate the economy's resources to those firms that can make the best use of those resources.*

However, there are other people who are skeptical about the claim that the price of a company's stock is closely tied to the company's efficiency. These observers believe that the demand for stock is disproportionately influenced by short-term developments in a company's profitability and that the market pays little attention to management decisions that promote the company's long-term earnings growth. These critics sometimes suggest that the stock market is close to a gambling casino in which hunch, rumor, and superstition have a critical influence on prices. (We will say more about this later in the chapter.)

THE RECENT SURGE IN TAKEOVER BATTLES

Small groups of individuals or enterprises sometimes buy enough stock in a company to gain control. When the purchaser is another firm, the process is referred to as a "merger" or an "acquisition." Periodically, there have been bursts of such takeover activity, with a very large rise in the number of mergers and acquisitions. Corporate "raiders" shook up the stock market and the managements of a number of corporations in the 1980s by trying to take over firms that they did not then control. This boom in takeovers slackened during the recession of the early 1990s, but has surged again recently. 1995 saw a record-breaking $458 billion worth of mergers and acquisitions, with such headline-making deals as IBM's $3.5 billion purchase of Lotus, and Seagram's $5.7 billion purchase of MCA. As this book was being written, takeover activity in the first quarter of 1996 had already approached nearly $110 billion.[3]

A **TAKEOVER** is the acquisition by an outside group (the raiders) of a controlling proportion of a company's stock. When the old management opposes the takeover attempt, it is called a *hostile takeover attempt.*

A **takeover** occurs when a group of outside financiers buys a sufficient amount of company stock to gain control of the firm. Often, the new controlling group will simply fire the current management and substitute a new chairman, president, and other top officers.

A company becomes a tempting target for a takeover attempt if the price of its stock is very low in comparison with the value of its plant, equipment, and other assets or when a company's earnings seem very low compared to their potential level. This may be because the firm's current management is believed not to be very competent or perhaps because the demand for a company's stocks is inordinately influenced by short-term developments, such as temporarily low profits.

An attempt to acquire the company by a group unfriendly to current management is called a *hostile takeover.* Naturally, the officers of the corporation will try to fight it off since they do not want to lose their high-paying jobs. They can fight back in many ways. For example, they can try to arrange instead for a "friendly takeover" by a group of investors whom they like better. They may also deliberately attempt to sabotage the company—often, by selling some of its most valuable parts in order to make what is left of the firm unattractive to the group attempting the takeover. Management may seek to bribe the takeover group to go away by offering a very high price for the stock that it already has managed to acquire. Indeed, takeovers are often attempted in the hope that the management will be forced to offer such a bribe (called *greenmail*) to those who threaten to take over the company.

In the second half of the 1980s, when a large number of takeover battles broke out, the issue received a good deal of publicity and set off a heated debate over its pros and cons. People who argue against strong legal restrictions on takeover activity point out that it is the most effective means to rid companies of incompetent managements and so helps to keep the economy at peak efficiency. They also argue that this activity helps "create stockholder value," that is, drive the price of an undervalued company's stock up to its true economic value.

But advocates of stricter regulations or inhibition of takeovers argue that stockholders who are innocent bystanders can be badly hurt in the process, as when management pays a large bribe to the takeover group or sells off a valuable part of the company when it should not be sold. Moreover, those seeking to buy a large percentage of the company's shares will try to do so as secretly

[3]"Merger Pace Races toward Another Record Year," C-Reuters@clari.net (Reuters), April 4, 1996.

as possible, hoping to obtain the stock cheaply. Critics claim that in the process, the raiders, in effect, cheat those who sold them the stock.

Opponents of takeovers also point out that the time taken by bright, talented people in planning and carrying out the strategies and counterstrategies uses up a valuable resource that could be better used elsewhere. On this view, takeover activity absorbs some of the nation's most capable individuals in financial manipulation rather than productive and innovative activity. These critics are wrong, however, when they argue that the billions of dollars that change hands in a takeover battle tie up the nation's capital wastefully or "use up" the economy's credit supply. Little or no capital (machinery, factories, and the like) is tied up in a takeover process. The money and credit that are used are simply transferred from one group of persons to another.

Often, takeovers have been financed by "junk bonds," that is, raiders issue bonds to raise the money which they need to acquire control of the target corporation. These bonds are frequently backed only by the profits that the raiders expect to grow out of their acquisition. Such profits may arise from more efficient management. The new owners may also sell off valuable portions of the corporation's activities (one of its successful products, for example) which they purchased cheaply because the corporation's stock price was low before the takeover.

Junk bonds backed only by such earnings prospects after the takeover are considered risky because of the danger that those promised profits may never materialize. A takeover financed in this way is called a "leveraged buyout" because the raider risks little of his or her own money in the process. Instead, the raider's limited resources are levered upward with the aid of other people's money—the money supplied by the junk bond purchasers. Critics of the process also note that it leaves the firm saddled with a heavy debt—its obligation to the junk bonds' purchasers.

THE ISSUE OF SPECULATION

Individuals who engage in **SPECULATION** deliberately invest in risky assets, hoping to obtain profits from future changes in the prices of these assets. (See the box on the next page for a description of a particularly risky speculative instrument, the *derivative*.)

Dealings in securities are often viewed with hostility and suspicion because they are thought to be an instrument of **speculation**. When something goes wrong in the market, say, when there is a sudden fall in prices, speculators are often blamed. The word *speculators* is used by editorial writers as a term of strong disapproval, implying that those who engage in the activity are parasites who produce no benefits for society and often do it considerable harm.

Economists disagree vehemently with this judgment. They say that speculators perform two vital economic functions:

1. Speculators sell *protection from risk* to other people, much as a fire insurance policy sells protection from risk to a homeowner.
2. They help to smooth out price fluctuations by purchasing items when they are abundant (and cheap) and holding them and reselling them when they are scarce (and expensive). In that way, speculators play a vital economic role in helping to alleviate and even prevent shortages.

Some examples from outside the securities markets will make the role of speculators clear. A ticket broker attends a preview of a new musical comedy and suspects that it is likely to be a hit. He decides to speculate by buying a large block of tickets for future performances. In that way, he takes over part of the producer's risk, for the producer now has some hard cash and has reduced her

HOW TO LOSE A BILLION DOLLARS IN ONE EASY STEP: THE DERIVATIVES CRAZE

A number of recent spectacular financial failures have been closely tied to a relatively new, and very risky, financial instrument: the derivative. First was the nearly $2 billion loss and resulting bankruptcy of Orange County, California; second was the collapse of a 233-year-old British investment bank, Barings PLC, after one of its employees (a 28-year-old "rogue" futures trader named Nicholas Leeson) "bet the ranch" on the movement of the Japanese stock market (and lost $1.46 billion).

Derivatives, one of the fastest growing areas of finance, are contracts so-named because they "derive" their value from the price movements of an underlying investment, such as a group of stocks, bonds, or commodities. Businesses buy these contracts in an effort to hedge or insure against sudden changes in interest rates or currency values. They also can be used to speculate in the markets.

The enormous financial losses that companies and other institutions have suffered as a result of their misguided investments in the very volatile derivatives market have brought calls for stringent regulation of these complex financial instruments. The Securities and Exchange Commission is developing new rules to require public companies to tell their stockholders the details of the company's derivatives dealings.



SOURCES: "The Dynamite and the Derivatives," *The Wall Street Journal*, February 28, 1995, p. A21; "Derivative Regulations Debated," *Clari.Net e.News* (http://clari.net), AP/Reuters, January 5, 1995; "Derivatives Changes Planned," Ibid., November 9, 1995; Richard W. Stevenson, "Big Gambles, Lost Bets Sank a Venerable Firm," *The New York Times*, March 3, 1995, pp. A1 and D15; "Broken Bank: Barings PLC Officials May Have Been Aware of Trader's Position," *The Wall Street Journal*, March 6, 1995, pp. A1 and A7; "Orange County Pleads for State Aid," *San Francisco Chronicle*, March 9, 1995.

inventory of risky tickets. If the show opens and is a flop, the broker will be stuck with the tickets. If it is a hit, he can sell them at a premium, if the law allows (and be denounced as a speculator or a "scalper").

Similarly, speculators enable farmers or producers of metals and other commodities whose future price is uncertain to get rid of their risk. A farmer who has planted a large crop but who fears its price may fall before harvest time can protect himself by signing a contract for future delivery at an agreed-upon price at which the speculator will purchase the crop when it comes in. In that case, if the price happens to fall, it is the speculator and not the farmer who will suffer the loss. Of course, if the price happens to rise, the speculator will reap the gain—that is the nature of risk bearing. The speculator who has agreed to buy the crop at the preset price, regardless of market conditions at the time the sale takes place, has, in effect, sold an insurance policy to the farmer. Surely this is a useful function.

The second role of speculators is perhaps even more important; in effect, they accumulate and store goods in periods of abundance and make goods available in periods of scarcity. Suppose that the speculator has reason to suspect that next year's crop of a storable commodity will not be nearly as abundant as this year's. She will buy some now, when it is cheap, for resale when it becomes scarce and expensive. In the process, she will smooth out the swing in prices by adding her purchases to the total market demand in the period of low prices (which tends to bring the price up), and bringing in her supplies during the period of high prices (which tends to push the price down).[4]

Thus, the successful speculator will help to relieve matters during periods of extreme shortage. There are cases in which they literally helped to relieve famine

[4]For a diagrammatic analysis of this role of speculation, see Review Question 7 at the end of the chapter.

by releasing the supplies they had deliberately hoarded for such an occasion. Of course, they are cursed for the high prices charged on such occasions. But those who curse them do not understand that prices might have been even higher if the speculator's foresight and avid pursuit of profit had not provided for the emergency. On the securities market, famine and severe shortages are not an issue, but the fact remains that successful speculators tend to reduce price fluctuations by increasing demand for stocks when prices are low and contributing to supply when prices are high.

Far from aggravating instability and fluctuations, speculators work as hard as they can to iron out fluctuations by buying when prices are low and selling when prices are high, for that is how they make their profits.

STOCK PRICES AS RANDOM WALKS

The beginning of this chapter cited evidence that the best professional securities analysts have a forecasting record so miserable that investors may do as well predicting earnings by hunch, superstition, or any purely random process as they would by following professional advice. (See the box on page 344–345 about predicting the stock market's performance via the outcome of the Super Bowl.)

Does this mean that analysts are incompetent people who do not know what they are doing? Not at all. Rather, there is fairly strong evidence that they have undertaken a task that is basically impossible.

How can this be so? The answer is that to make a good forecast of any variable—GDP, population, or fuel usage—there must be something in the past whose behavior is closely related to the future behavior of the variable whose path we wish to predict. If a 10 percent rise in this year's consumption always produces a 5 percent rise in next year's GDP, this fact can help us predict future GDP on the basis of current observations. But if we want to forecast the future of a variable whose behavior is completely unrelated to the behavior of *any* current or past variable, there is no objective evidence that can help us make that forecast. Throwing darts or gazing into a crystal ball are no less effective than analysts' calculations.

There is a mass of statistical evidence that the behavior of stock prices is largely unpredictable. In other words, the behavior of stock prices is essentially random; the paths they follow are what statisticians call **random walks.** A random walk is like the path followed by a drunk. All we know about his position after his next step is that it will be given by his current position plus whatever random direction his next haphazard step will carry him. The relevant feature of randomness, for our purposes, is that it is by nature unpredictable, which is just what the word *random* means.

The time path of a variable such as the price of a stock is said to constitute a **RANDOM WALK** if its magnitude in one period (say, May 2, 1997) is equal to its value in the preceding period (May 1, 1997) plus a completely random number. That is: Price on May 2, 1997 = Price on May 1, 1997 + Random number, where the random number (positive or negative) might be obtained by a roll of dice or some such procedure.

If the evidence that stock prices approximate a random walk stands up to research in the future as it has so far, it is easy enough to understand why stock market predictions are as poor as they are. Analysts are trying to forecast behavior that is basically random; in effect, they are trying to predict the unpredictable.

Two questions remain. First, does the evidence that stock prices follow a random walk mean that investment in stocks is a pure gamble and never worthwhile? Second, how does one explain the random behavior of stock prices?

To answer the first question, it is false to conclude that investment in stocks is generally not worthwhile. The statistical evidence is that, over the long run, stock prices *as a whole* have had a fairly marked upward trend, perhaps reflect-

FOOTBALL AND FINANCIAL FORECASTING

The following excerpt from a column in the business section of *The New York Times* suggests some of the gimmicks that stock market analysts turn to in a desperate effort to predict stock prices.

The Market Predicts the Super Bowl and Vice Versa, So They Say

The bull market is good for another year, no matter which team wins this year's Super Bowl. And that is just as well, because the stock market right now sees the game as very close, although last week's sharp fall in share prices has given the Dallas Cowboys a small edge.

Those are the results of two theories relating the stock market to the Super Bowl. It is, of course, absurd to think there is any relation between the two, but that doesn't stop Wall Street from paying attention.

First, the future of the market. That theory has been around for years, occasionally tweaked to make it work better. In its current form, it says that if a team from the old National Football League wins the game, the stock market will rise during the year until the next game. But if a team from the old American Football League wins, share prices will fall.

The good news for the bulls was locked in a week ago: All four teams in the conference championship game played Sunday, including the two from the American Football Conference, were old N.F.L. teams. So the bulls clearly have it.

That theory has worked this year. A year ago, it was San Francisco (old N.F.L.) over San Diego (old A.F.L.) in the Super Bowl, which the theory said meant prices would rise. Sure enough, the Dow Jones industrial average is now 31 percent over the 3,857.99 level where it stood entering that game. That brings the record for the indicator to 25 of 29. The failures came after the 1970, 1987, 1990 and 1994 games. In 1970, the game falsely forecast falling share prices; in the other years it was wrongly bullish.

Dallas was the winner in the last of those games, but the market did go up after its other three victories. Similarly, share prices rose after each of Pittsburgh's four previous victories.

The newer theory, written about in this column seven years ago, attempts to use the stock market to forecast the game. It holds that if the Dow rises from the end of November until game day, the franchise from later in the alphabet will win.

In this case, that means the Pittsburgh Steelers will benefit from a bull market, while the Dallas Cowboys will be favored if share prices fall.

The Dow Jones industrial average ended November at 5,074.49, and subsequently got up to a high of 5,216.47. But last week's sharp fall, the worst week for the index in more than a year, left the average at 5,061.12, below the magic number, and it fell further, to 5,043.78, yesterday. But there are nine trading

ing the long-term growth of the economy. Thus, the random walk does not proceed in just any direction—rather, it represents a set of erratic movements *around a basic upward trend in stock prices.*

Moreover, it is not in the *overall* level of stock prices that the most pertinent random walk occurs, but in the performance of one company's stock compared with another's. For this reason, professional advice may be able to predict that investment in the stock market is likely to be a good thing over the long haul. But, if the random walk evidence is valid, there is no way professionals can tell us which of the available stocks is most likely to go up—that is, which combination of stocks is best for the investor to buy.

The only appropriate answer to the second question is that no one is sure of the explanation. There are two widely offered hypotheses—each virtually the opposite of the other. The first asserts that stock prices are random because clever professional speculators are able to foresee almost perfectly every influence that is *not* random. For example, suppose that a change occurs that makes the probable earnings of some company higher than had previously been expected. Then, according to this view, the professionals will instantly become aware of this change and immediately buy enough to raise the price of the stock accordingly. Then, the only thing for that stock price to do between this year and next is wander randomly, because the professionals cannot predict random movements, and hence they cannot force current stock prices to anticipate them.

The other explanation of random behavior of stock prices is at the opposite pole from the view that all nonrandom movements are wiped out by supersmart professionals. This view holds that people who buy and sell stocks have learned that they cannot predict future stock prices. As a result, they react to any signal,

days left before the game, with a gain of just over 30 points needed to forecast a Pittsburgh win.

When the theory regarding the Dow forecasting the Super Bowl game was first written about in this column, it had correctly forecast 12 of the previous 14 games. Since then, it has gotten 6 of 7 correct, bringing its total record to 18 of 21.

By contrast, the Las Vegas point spread has correctly picked the winner in 16 of 20 games over the same stretch, with one game, in 1982, going off with no favorite.

The three games in which the football theory has failed to forecast the Super Bowl game were in 1982, when San Francisco beat Cincinnati; in 1986, when Chicago topped New England; and in 1990, when San Francisco beat Denver.

People who buy or sell stocks or place bets on football games based on such indicators probably should not be allowed to manage their own money. But the existence of indicators with such apparently superb records nonetheless does serve a use, by reminding investors that the fact that two things are correlated may or may not indicate that they are related. That is useful to consider when looking at other indicators that seem to make more sense.

But if you are unwilling to accept that warning, here's another correlation to ponder. The last three times the economy had at least one quarter in which it did not show growth, after adjusting for inflation, were in 1982, 1986 and 1990. And, as it happens, they were also the years in which the stock market failed to forecast the Super Bowl.

Therefore, if this indicator does not get the game right on January 28, you should watch out for a recession this year.

SOURCE: Floyd Norris, "The Market Predicts the Super Bowl," *The New York Times*, January 16, 1996, p. D12.



however irrational and irrelevant it appears. If the president catches cold, stock prices fall. If an astronaut's venture is successful, prices go up. According to this view, investors are, in the last analysis, trying to predict not the prospects of the economy or of the company whose shares they buy, but the supply and demand behavior of other investors, which will ultimately determine the course of stock prices. Since all investors are equally in the dark, their groping can only result in the randomness that we observe. The classic statement of this view of stock market behavior was provided by Lord Keynes, a successful professional speculator himself:

> Professional investment may be likened to those newspaper competitions in which the competitors have to pick out the six prettiest faces from a hundred photographs, the prize being awarded to the competitor whose choice most nearly corresponds to the average preferences of the competitors as a whole; so that each competitor has to pick not those faces which he himself finds prettiest, but those which he thinks likeliest to catch the fancy of the other competitors, all of whom are looking at the problem from the same point of view. It is not a case of choosing those which, to the best of one's judgment, are really the prettiest, nor even those which average opinion genuinely thinks the prettiest. We have reached the third degree where we devote our intelligences to anticipating what average opinion expects the average opinion to be. And there are some, I believe, who practice the fourth, fifth and higher degrees.[5]

[5]John Maynard Keynes, *The General Theory of Employment, Interest, and Money* (New York: Harcourt Brace, 1936), p. 156.

This may help to explain the impressive rise of the stock market from a Dow Jones index of 800 in 1982 to 2,700 in 1987, its 700-point fall in two consecutive trading days in October 1987, and its several sharp fluctuations followed by an upward trend since then.

SUMMARY

1. The three basic types of firms are **corporations, partnerships,** and **individual proprietorships.** Most U.S. firms are individual proprietorships, but most U.S. manufactured goods are produced by corporations.
2. Individual proprietorships and partnerships have tax advantages over corporations. But corporate investors have greater protection from risk because they have **limited liability**—they cannot be asked to pay more than they have invested in the firm.
3. Higher taxation of corporate earnings tends to limit the things in which corporations can invest and may lead to inefficiency in resource allocation.
4. A **common stock** is a share in the ownership of a company. A **bond** is an I.O.U. by a company for money lent to it by the bondholder. Many observers argue that the purchase of a stock also really amounts to a loan to the company—a loan that is riskier than the purchase of a bond.
5. If interest rates rise, bond prices will fall. In other words, if some bond amounts to a contract to pay 8 percent and the market interest rate goes up to 10 percent, people will no longer be willing to pay the old price for that bond.
6. Corporations finance their activities mostly by **plowback** (that is, by retaining part of their earnings and putting it back into the company) or by sales of stocks and bonds.
7. If stock prices correctly reflect the future prospects of different companies, it is easier for promising firms to raise money because they are able to sell each stock they issue at favorable prices.
8. Bonds are relatively risky for the firms that issue them, but they are fairly safe for their buyers, because they are a commitment by those firms to pay fixed annual amounts to the bondholders whether or not the companies make money that year. But stocks, which do not promise any fixed payment, are relatively safe for the companies and risky for their owners.
9. A portfolio is a collection of stocks, bonds, and other assets of a single owner. The greater the number and variety of securities and other assets it contains, the less risky it is.
10. A corporation is said to be taken over when an outside group buys enough stock to get control of the firm's decisions. **Takeovers** are a useful way to get rid of incompetent management or to force management to be efficient. However, the process is costly and leads to wasteful defensive and offensive activities.
11. **Speculation** affects stock market prices, but (contrary to what is widely assumed) there is reason to believe that speculation actually *reduces* the frequency and sizes of price fluctuations. Speculators are also useful to the economy because they undertake risks that others wish to avoid, thereby, in effect, providing others with insurance against risk.
12. Statistical evidence indicates that individual stock prices behave randomly.

KEY TERMS

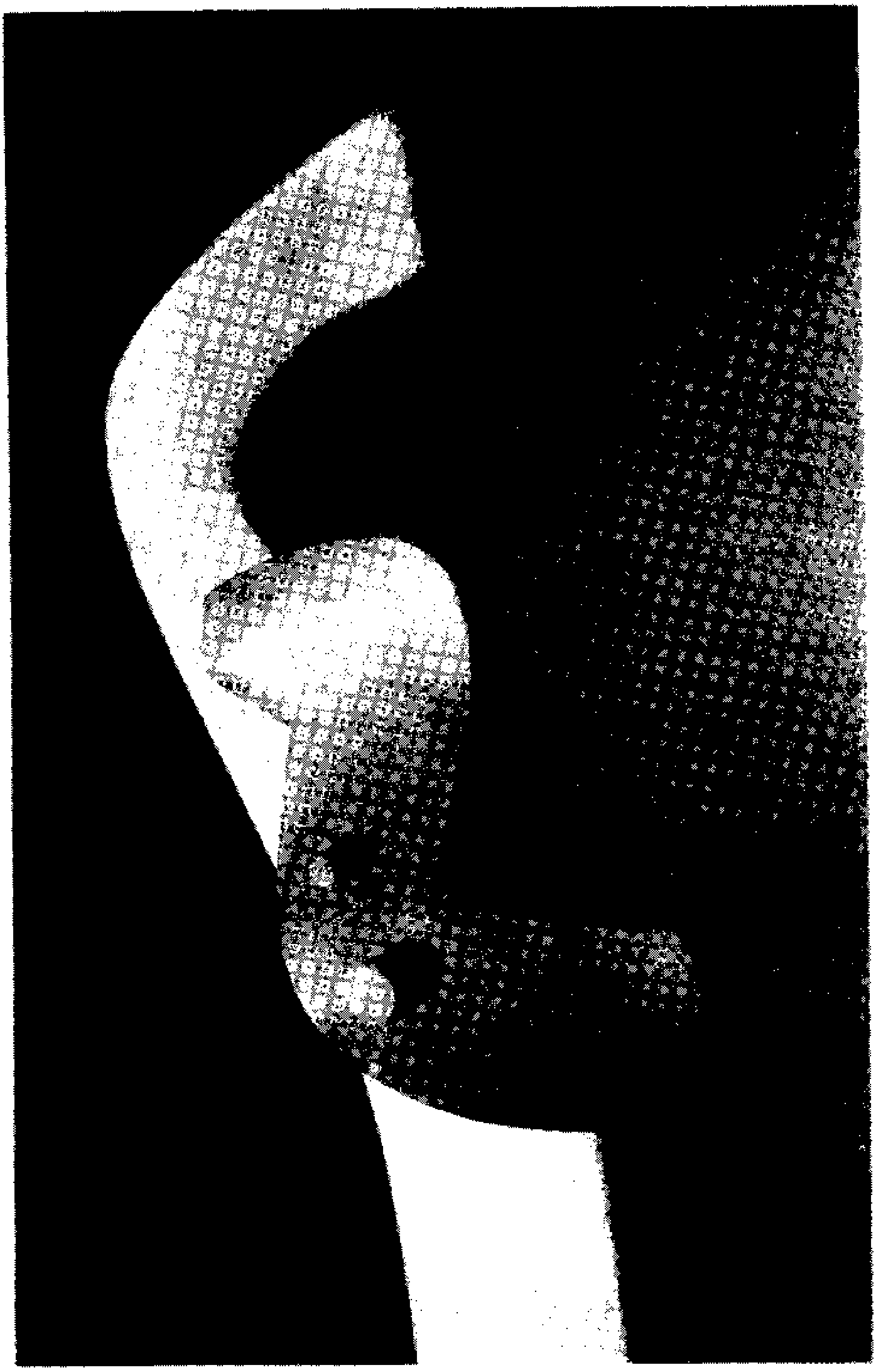
Proprietorship
Unlimited liability
Partnership
Corporation
Limited liability

Double taxation
Common stock
Bond
Plowback or retained earnings
Portfolio diversification

Stock exchanges
Takeovers
Speculation
Random walk

QUESTIONS FOR REVIEW

1. Why would it be difficult to run General Motors as a partnership or an individual proprietorship?
2. Do you think it is fair to tax a corporation more than a partnership doing the same amount of business? Why or why not?
3. If you hold shares in a corporation and management decides to plow back the company's earnings some year instead of paying dividends, what are the advantages and disadvantages to you?
4. Suppose that interest rates are 6 percent in the economy and a safe bond promises to pay \$3 a year in interest forever. What do you think the price of the bond will be? Why?
5. Suppose in the economy in the previous example, interest rates suddenly fall to 3 percent. What will happen to the price of the bond that pays \$3 per year?
6. If you want to buy a stock, when might it be to your advantage to buy it using a market order? When will it pay to use a limit order?
7. Show in diagrams that if a speculator were to buy when price is high and sell when price is low, he would increase price fluctuations. Why would it be in his best interest *not* to do so? (*Hint:* Draw two supply-demand diagrams, one for the high-price period and one for the low-price period. How would the speculator's activities affect these diagrams?)
8. If stock prices really are a random walk, can you nevertheless think of good reasons for getting professional advice before investing?
9. Hostile takeovers often end up in court when managements attempt to block them and raiders accuse the managements of selfishly sacrificing the interests of stockholders. The courts often look askance at "coercive" offers by raiders—an offer to buy, say, 20 percent of the company's stock by a certain date from the first stockholders who offer to sell. By contrast, they take a more favorable attitude toward "noncoercive" offers to buy any and all stock supplied at announced prices. Do you think the courts are right to reject "coercive offers" but prevent management from blocking "noncoercive" offers? Why?
10. In "program trading," computers decide when to buy or sell stocks on behalf of large, institutional investors. The computers then carry out those transactions with electronic speed. Critics claim that this is a major reason why stock prices rose and fell sharply in the 1980s. Is this plausible? What other influences may have been important?



PART IV

THE DISTRIBUTION OF INCOME

CHAPTER 15

Pricing the Factors of Production

Rent is that portion of the produce of the earth which is paid to the landlord for use of the original and indestructible powers of the soil.
David Ricardo

Chapter 13 mentioned that the market mechanism cannot be counted on to distribute income in accord with ethical notions of "fairness" or "justice," and listed this as one of the market's shortcomings. But there is much more to say about how income is distributed in a market economy and about how governments interfere with this process. These are the subjects of Part IV of this book.

All of us are familiar with the broad outlines of how the market mechanism distributes income. Everyone owns some *factors of production*—the inputs used in the production process. Many of us have only our own labor; but some of us also have funds that we can lend, land that we can rent, or natural resources that we can sell. We sell these factors on markets at prices determined by supply and demand. So the distribution of income in a market economy is determined by the level of employment of the factors of production and by their prices. For example, if wages are rather high and are fairly equal among workers, and if unemployment is low, then few people will be poor. But if wages are low and unequal and unemployment is high, then many people will be poor.

ENTREPRENEURSHIP is the act of starting new firms, introducing new products and technological innovations, and, in general, taking the risks that are necessary in seeking out business opportunities.

For purposes of discussion, we group the factors of production into five broad categories: land, labor, capital, exhaustible natural resources, and a rather mysterious input called **entrepreneurship.** Labor is important enough to merit a full chapter (which follows this one), and we will discuss exhaustible natural resources in the second half of Chapter 21. Here we will study the payments made for the use of the other three factors: the interest paid to capital, the rent of land, and the profits earned by entrepreneurs.

But since there are so many misperceptions about the distribution of income among capitalists, landlords, and entrepreneurs, we will first look briefly at how much these three groups earn in reality. According to U.S. data for 1994, interest payments accounted for about 7.5 percent of national income; land rents were minuscule, accounting for half of 1 percent; corporate profits accounted for about 10

percent and income of other proprietors for about 9 percent.[1] In total, the returns to all the factors of production that we deal with in this chapter amounted to about one-quarter of national income. Where did the rest of it go? The answer is that almost three-quarters of national income was composed of employee compensation—wages and salaries. The huge share of labor in national income is one of the reasons why the next chapter is devoted entirely to this subject.

In this chapter we will encounter examples of serious misunderstandings about the nature of distribution and about what can be done to influence it. That's because the distribution of income is the one area in economics in which any one individual's interests almost inevitably conflict with someone else's. By definition, if a larger share of the total income is distributed to me, a smaller share will be left for you. It is also a topic about which emotions run high, and the facts or the logic of the issues are often ignored.

THE PRINCIPLE OF MARGINAL PRODUCTIVITY

By now it will not surprise you to learn that factor prices are analyzed in terms of supply and demand. The supply sides of the markets for the various factors differ enormously from one another, which is why we must consider each factor market separately. But we can use one basic principle, the *principle of marginal productivity,* to explain the *demand* for any input by a profit-maximizing firm, given the price of that input. Before reviewing the principle, it will be useful to recall two concepts introduced in Chapter 7: **marginal physical product (MPP)** and **marginal revenue product (MRP).**[2]

The **MARGINAL PHYSICAL PRODUCT (MPP)** of an input is the increase in output that results from a 1-unit increase in the use of the input, holding the amounts of all other inputs constant.

The **MARGINAL REVENUE PRODUCT (MRP)** of an input is the money value of the additional sales that a firm obtains by selling the marginal physical product of that input.

Table 15-1 helps us review these two concepts by recalling the example of Florence Farmer, who had to decide how much corn to feed her chickens. The marginal *physical* product (MPP) column tells us how many additional pounds of chicken each additional bag of corn will yield. For example, according to the table, the fourth bag increases output by 34 pounds. The marginal *revenue* product (MRP) column tells us how many dollars this marginal physical product is worth. In the example in the table, we assume chicken always sells at $0.75 per pound, so the marginal revenue product of the fourth bag of corn is $0.75 per pound times 34 pounds, or $25.50 (last column of the table).

The marginal productivity principle states that when factor markets are competitive, it always pays a profit-maximizing firm to hire the quantity of any input that makes the marginal revenue product equal to the price of the input.

The basic logic behind the principle is both simple and powerful. We know that the firm's profit from acquiring an additional unit of an input is the input's marginal revenue product minus the marginal cost (the price of the additional unit of input). Thus, if the input's marginal revenue product is, for example,

[1] U.S. Bureau of the Census, *Statistical Abstract of the United States 1995,* 115th ed. (Washington, D.C.: U.S. Government Printing Office, September 1995).

[2] To review these concepts, see Chapter 7, pages 149–152.

TABLE 15-1 FLO'S SCHEDULES FOR TPP, MPP, APP, AND MRP OF CORN

Input (bags of corn)	TPP: Total Physical Product (chicken output in pounds)	MPP: Marginal Physical Product (pounds per bag)	APP: Average Physical Product (pounds per bag)	MRP: Marginal Revenue Product (dollars per bag)
0	0.0			
1	14.0	14.0	14.0	10.50
2	36.0	22.0	18.0	16.50
3	66.0	30.0	22.0	22.50
4	100.0	34.0	25.0	25.50
5	130.0	30.0	26.0	22.50
6	156.0	26.0	26.0	19.50
7	175.0	19.0	25.0	14.25
8	184.0	9.0	23.0	6.75
9	185.4	1.4	20.6	1.05
10	180.0	−5.4	18.0	−4.05
11	165.0	−15.0	15.0	−11.25
12	144.0	−21.0	12.0	−15.75

greater than its price, it will pay the profit-seeking firm to hire more of it because an additional unit of input brings the firm an addition to revenue that exceeds its cost. The firm should increase the quantity purchased up to the amount at which diminishing returns reduce the MRP to the level of the input's price. By similar reasoning, if MRP is less than price, then the firm is using too much of the input. We see in Table 15-1 that about seven bags is the optimal amount of corn for Flo to use each week, because an eighth bag brings in a marginal revenue product of only $6.75, which is less than the $10.00 cost of buying the bag.

One corollary of the principle of marginal productivity is obvious: The quantity of the input demanded depends on its price. The lower the price of corn, the more it pays the farm to buy. In our example, it pays Florence to use between seven and eight bags when the price per bag is $10.00. But if corn were more expensive, say, $20.00 per bag, that high price would exceed the value of the marginal product of either the sixth or seventh bag. It would, therefore, pay the firm to stop at five bags of corn. Thus, *marginal productivity analysis shows that the quantity demanded of an input normally declines as the price of the input rises.* The "law" of demand applies to inputs just as it applies to consumer goods.

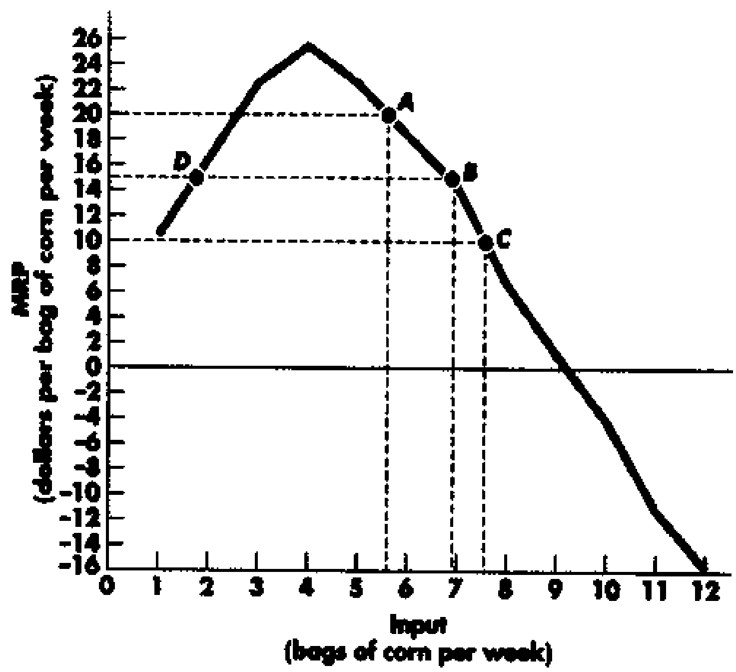
THE DERIVED DEMAND CURVE FOR AN INPUT

We can, in fact, be much more specific than this, for the marginal productivity principle tells us precisely how the demand curve for any input is derived from its marginal revenue product (MRP) curve.

Figure 15-1 presents graphically the MRP schedule from Table 15-1 for three different possible prices for a bag of corn: $20.00, $15.00, and $10.00. At a price

FIGURE 15-1 A MARGINAL REVENUE PRODUCT SCHEDULE

This diagram depicts the data in Table 15-1, which show how the marginal revenue product (MRP) of corn first rises and then declines as more and more corn is used. Since the optimal purchase rule is to keep increasing the use of corn until its MRP is reduced to the price of corn, the downward-sloping portion of the MRP curve is Florence Farmer's demand curve for corn.



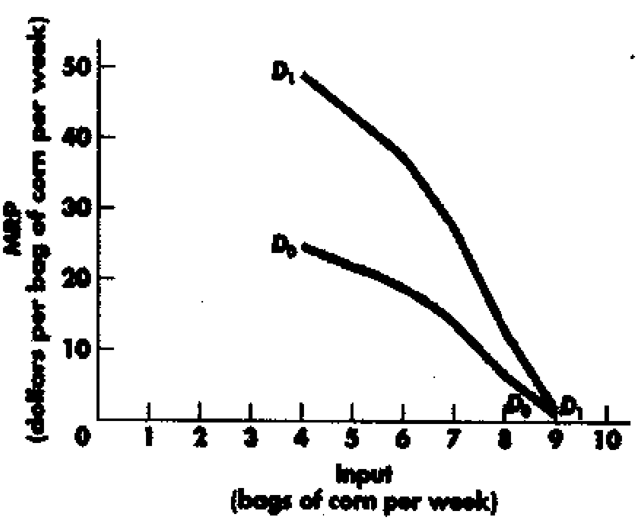
of $20.00 per bag, we see that the quantity demanded is about 5.6 bags of corn per week (point *A*) because at that point MRP equals price. Similarly, if the price of corn drops to $15.00 per bag, quantity demanded rises to about 6.8 bags per week (point *B*). Finally, should the price fall all the way to $10.00 per bag, the quantity demanded would be about 7.5 bags per week (point *C*). Points *A*, *B*, and *C* are therefore three points on the demand curve for corn. By repeating this exercise for any other price, we learn that:

The demand curve for any input is the downward-sloping portion of its marginal revenue product curve.

Why is the demand curve restricted to only the *downward-sloping portion* of the MRP curve? The logic of the marginal productivity principle dictates this. For example, if the price of corn were $15.00 per bag, Figure 15-1 shows that there are two input quantities for which MRP = *P*: (approximately) 2 bags (point *D*) and 6.8 bags (point *B*). But point *D* cannot be the optimal stopping point because the MRP of a third bag ($22.50) is greater than the cost of the third bag ($15.00), so that the firm makes more money by expanding its input use beyond 2 bags per week. A similar profitable opportunity for expansion occurs any time the MRP curve slopes upward at the current price, since that means that an increase in the quantity of input used by the firm will raise MRP above the input's price. It follows that a profit-maximizing firm will always demand an input quantity that is in the range where MRP is diminishing.

The demand for corn or labor (or for any other input) is called a *derived demand* because it is derived from the underlying demand for the final product (chicken in this case). For example, suppose that a surge in demand drove the price of chicken to $1.50 per pound. Then, at each level of corn usage, the marginal revenue product would be twice as large as when chicken fetched only $0.75 per

FIGURE 15-2 A SHIFT IN THE DEMAND CURVE FOR CORN



If the price of chicken goes up, the marginal *revenue* product curve of corn shifts upward—from D_0D_0 to D_1D_1 in the diagram—even though the marginal *physical* product curve has not changed. In this sense *a greater demand for chicken leads to a greater derived demand for corn.*

pound. This is shown in Figure 15-2 as an upward shift of the (derived) demand curve for corn, from D_0D_0 to D_1D_1.[3] We conclude that, in general:

An outward shift in the demand curve for any commodity causes an outward shift of the derived demand curve for all factors utilized in the production of that commodity.

Conversely, an inward shift in the demand curve for a commodity leads to inward shifts in the demand curves for factors used in producing that commodity.

This completes our discussion of the *demand* side of the analysis of input pricing. The most noteworthy feature of the discussion is the fact that the same marginal productivity principle serves as the foundation for the demand schedule for each and every type of input. In particular, as we will see in the next chapter, the marginal productivity principle serves as the basis for the determination of the demand for labor—that crucial input whose financial reward plays so important a role in an economy's standard of living. On the demand side, one analysis fits all.

Things are very different when we turn to the supply side, however. Here we must deal with each of the main factors of production individually because each involves a somewhat different story. We must do this in order to see how their earnings are determined by the interaction of demand *and* supply. We begin with *interest payments*, the return on loans of money. First, we must define a few key terms.

INVESTMENT, CAPITAL, AND INTEREST

The rate of interest is the price at which funds can be rented (borrowed). And, just like other factor prices, interest rates are determined by supply and demand. There are many ways in which funds are loaned (that is, rented to users): home

[3] To make the diagram easier to read, the (irrelevant) upward-sloping portion of each curve has been omitted.

mortgages, corporate or government bonds, consumer credit, and so on. On the demand side of these credit markets are borrowers—people or institutions that, for one reason or another, wish to spend more than they currently have. In business, loans are used primarily to finance investment. To the business executive who borrows funds in order to finance an **investment** and pays interest in return, the funds really represent an intermediate step toward the acquisition of the machines, buildings, inventories, and other forms of physical **capital** that the firm will purchase.

INVESTMENT is the *flow* of resources into the production of new capital. It is the labor, steel, and other inputs devoted to the *construction* of factories, warehouses, railroads, and other pieces of capital during some period of time.

CAPITAL refers to an inventory (a *stock*) of plant, equipment, and other productive resources held by a business firm, an individual, or some other organization.

Though the words *investment* and *capital* are often used interchangeably in everyday parlance, the distinction is important. Economists define investment as the rate at which capital *grows*. The higher the level of investment, the greater the rate of growth of the amount of capital in the possession of the investor. The relation between investment and capital can best be described by the analogy of filling a bathtub: The accumulated water in the tub is analogous to the *stock* of capital, while the flow of water from the faucet (which adds to the tub's water) is like the *flow* of investment. Just as the faucet must be turned on in order for more water to accumulate, the capital stock increases only when there is investment. If investment ceases, the capital stock stops growing. Notice that when investment is *zero*, the capital stock *remains constant*; it does not fall to zero any more than a bathtub suddenly becomes empty when you turn off the faucet.

The process of building up capital by investing and then using this capital in production can be divided into five steps. They are listed below and summed up in Figure 15-3.

Step 1. The firm decides to enlarge its stock of capital.

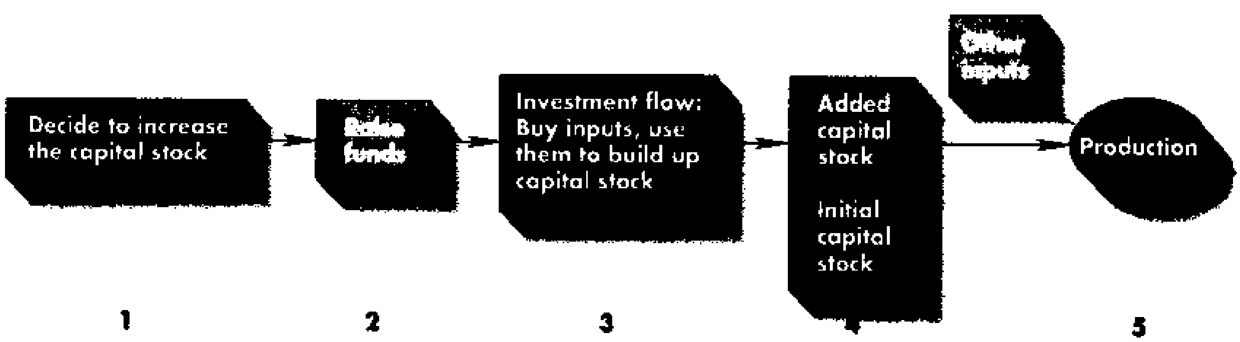
Step 2. It raises the funds to finance its expansion either from outside sources such as banks, or by holding on to part of its earnings rather than paying them out to the owners of the company.

Step 3. It uses these funds to hire the inputs to build factories, warehouses, and the like. This step is the act of *investment*.

Step 4. After the investment is completed, the firm ends up with a larger stock of *capital*.

Step 5. The capital is used (along with other inputs) either to expand production or to reduce costs. At this point, the firm starts earning *returns* on its investment.

FIGURE 15-3 THE INVESTMENT PRODUCTION PROCESS



The investor (1) decides to increase his capital stock, (2) raises funds, (3) uses the funds to buy inputs that produce capital stock like machinery and factory buildings (this step is called *investment*), (4) now holds more capital than before, and (5) uses this capital and other inputs to produce goods and services.

INTEREST is the payment for the use of funds employed in the production of capital; it is measured as a percentage per year of the value of the funds tied up in the capital.

Notice that what the investors put into the investment process is *money*, either their own or funds they borrow from others. Then, through a series of steps, the funds are transformed into a physical input suitable for use in production. If investors have borrowed the funds, they will someday return them to the lender with some payment for their use. This payment is called **interest**, and it is calculated as a percentage per year of the amount borrowed. For example, if an investor borrows $1,000 at an interest rate of 12 percent per year, the annual interest payment is $120.

The marginal productivity principle governs the quantity of funds demanded just as it governs the quantity of corn demanded for chicken feed. Specifically:

Firms will demand the quantity of borrowed funds that makes the marginal revenue product of the investment financed by the funds just equal to the interest payment charged for borrowing.

There is one noteworthy feature of capital that distinguishes it from other inputs, like corn, for example. When Florence Farmer feeds corn to her chickens, it is used once and then it is gone. But a blast furnace, which is part of a steel company's capital, normally lasts many years. The furnace is a *durable* good; and because it is durable, it contributes not only to today's production, but also to future production. This fact makes calculating the marginal revenue product more complex for a capital good than for other inputs.

To determine whether the MRP of a capital good is greater than the cost of financing it (that is, to decide whether an investment is profitable), we need a way to compare money values received at different times. To make such comparisons, economists and business people use a calculation procedure called *discounting*. We will explain discounting in detail in the appendix to this chapter, but it is not important that you master this technique in an introductory course. There are really only two important points to learn:

1. A sum of money received at a future date is worth less than the same sum of money received today.
2. This difference in values between money today and money in the future is greater when the rate of interest is higher.

It is not difficult to understand why this is so. Consider what you could do with a dollar that you received today rather than a year from today. If the annual rate of interest were 10 percent, you could lend it out (for example, by putting it in a bank account), and receive $1.10 in a year's time—your original $1.00 plus 10 cents interest. For this reason, money received today is worth more than the same number of dollars received later. Specifically, at a rate of interest of 10 percent per year, $1.10 to be received a year from today is equivalent to $1.00 of today's money. This illustrates the first of our two points.

Now suppose the annual rate of interest was 15 percent instead. In this case $1.00 invested today would grow to $1.15 (rather than $1.10) in a year's time, which means that $1.15 received a year from today would be equivalent to $1.00 received today, and so $1.10 a year in the future must now be worth less than $1.00 today. This illustrates the second point.

The rate of interest is a crucial determinant of the economy's level of investment, that is, of the amount of current consumption that consumers will choose to forgo in order to use the resources to build machines and factories that can increase the output of consumers' goods in the future. For that reason, the interest rate is crucial in determining the allocation of society's resources between

present and future—an issue that we discussed in Chapter 13 (pages 311–313). Let us see, then, how the magnitude of the interest rate is set by the market.

THE DOWNWARD-SLOPING DEMAND CURVE FOR FUNDS

The two attributes of discounting discussed above are all we need to explain why the quantity of funds demanded declines when the interest rate rises, that is, why the demand curve for funds has a negative slope.

Remember that the demand for borrowed funds is a *derived demand*, derived from the desire to invest in capital goods. But part, and perhaps all, of the marginal revenue product of a machine or a factory is received in the future. Hence, the value of the MRP *in terms of today's money* shrinks as the rate of interest rises. Why? Because future returns must be *discounted more* when the rate of interest rises, for reasons just discussed. The consequence of this shrinkage is that a machine that appears to be a good investment when the rate of interest is 10 percent may look like a terrible investment when the rate of interest is 15 percent. That is, the higher the discount rate, the fewer machines a firm will demand. Thus, the demand curve for machines and other forms of capital will have a negative slope—the higher the interest rate, the smaller the quantity that firms will demand.

As the rate of interest on borrowing rises, more and more investments that previously looked profitable start to look unprofitable. The demand for borrowing for investment purposes, therefore, is lower at higher rates of interest.

It should be noted that while this analysis clearly applies to a firm's purchase of capital goods such as plant and equipment, it can also apply to the company's purchases of land and labor. Both of these are often financed by borrowed funds, and the marginal products of these inputs may accrue only months or even years after the inputs have been bought and put to work. (For example, it may take quite some time before newly acquired agricultural land will yield a marketable crop.) For both reasons, then, a rise in the rate of interest will reduce the quantity demanded of investment goods like land and labor, just as it cuts the derived demand for investment in plant and equipment.

Figure 15-4 depicts a derived demand schedule for loans. Its negative slope illustrates the conclusion we have just stated:

The higher the interest rate, the less people and firms will want to borrow to finance their investments.

THE SUPPLY OF FUNDS

Somewhat different relationships arise on the supply side of the market for funds—where the *lenders* are consumers, banks, and other types of business firms. Funds lent out are usually returned to the owner (with interest) only over a period of time. Loans will look better to lenders when they bear higher interest rates, so it is natural to think of the supply schedule for loans as being upward sloping—at higher rates of interest, lenders supply more funds. Such a supply schedule is shown by the curve *SS* in Figure 15-5, where we also reproduce the demand curve, *DD*, from Figure 15-4.

However, not all supply curves for funds slope uphill to the right like curve *SS*. Suppose, for example, that Melinda Martinez is saving to buy a $10,000 boat in 3 years, and that if she lends money out at interest in the interim, she must

FIGURE 15-4 THE DERIVED DEMAND CURVE FOR LOANS

The rate of interest is the cost of a loan to the borrower. The lower the rate of interest, the more it will pay a business firm to borrow in order to finance new plant and equipment. That is why this demand curve has a negative slope.

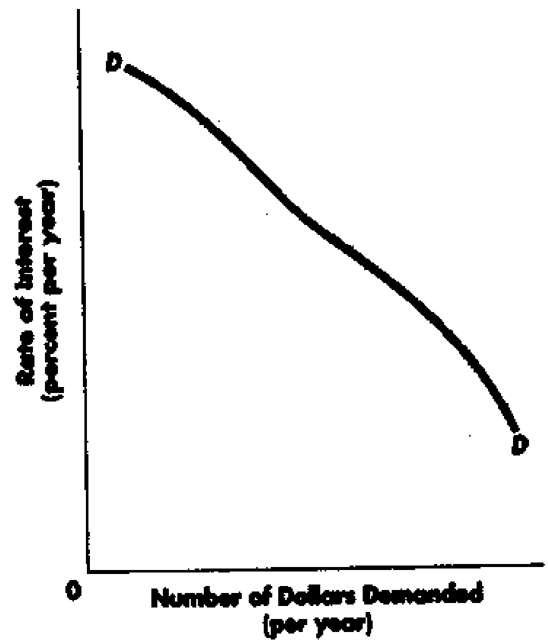
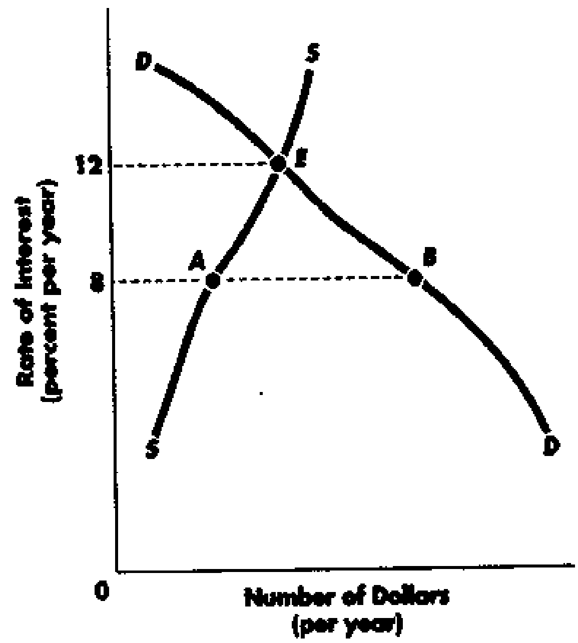


FIGURE 15-5 EQUILIBRIUM IN THE MARKET FOR LOANS

Here the free-market interest rate is 12 percent. At this interest rate, the quantity of loans supplied is equal to the quantity demanded. However, if an interest-rate ceiling is imposed, say, at 8 percent, the quantity of funds supplied (point A) will be smaller than the quantity demanded (point B).



save $3,000 a year to reach her goal. If interest rates were higher, she could save less than $3,000 each year and still reach her $10,000 goal. (The higher interest payments would, of course, contribute the difference.) So her saving (and lending) might decline. This argument only applies fully to savers, like Martinez, with a fixed accumulation goal. But similar considerations affect the calculations of other savers.

Generally, we do expect the quantity of loans supplied to rise at least somewhat when the interest reward rises, so the supply curve will have a positive slope, like *SS* in Figure 15-5. However, for reasons indicated in the previous paragraph, the increase in the economy's saving that results from a rise in the interest rate is usually quite small. That is why we have drawn the supply curve so steep. The rise in the amount supplied by some lenders is partially offset by the decline in the savings of Ms. Martinez who is putting money away to buy a boat, or Mr. Smith who is saving for his children's college tuition.

Having examined the relevant demand and supply curves, we are now in a position to examine the determination of the equilibrium rate of interest. This is summed up in Figure 15-5, in which the equilibrium is, as always, at point *E*, where quantity supplied and quantity demanded are equal. We conclude that the equilibrium interest rate on loans is 12 percent in the example in the graph.

THE ISSUE OF USURY LAWS: ARE INTEREST RATES TOO HIGH?

People have often been dissatisfied with the workings of the market in the determination of interest rates. Fears that interest rates, if left unregulated, would climb to exorbitant levels have made usury laws (which place upper limits on money-lending rates) quite popular in many times and places. Attempts to control interest payments date back to biblical days, and in the Middle Ages the influence of the church even led to total prohibition of interest payments in much of Europe. In the recent past usury laws governed maximum rates on consumer loans, home mortgages, and the like in some parts of the United States.

Unscrupulous lenders often manage to evade usury laws by charging interest rates even higher than the free-market equilibrium rate—including generous compensation to themselves for the risk of being caught at charging the illegal rates. But even when usury laws are effective, they interfere with the operation of supply and demand and, as we will demonstrate, they often harm economic efficiency.

Let us refer once again to Figure 15-5, but this time we assume it depicts the supply of loans by banks to consumers. Consider what happens if there is a usury law that prohibits interest of more than 8 percent per year on consumer loans. At this interest rate, the quantity supplied (point *A* in Figure 15-5) falls short of the quantity demanded (point *B*). This means that many applicants for consumer loans are being turned down even though the banks consider them to be credit-worthy.

Who generally gains and who loses from this usury law? The gainers are easiest to identify: those lucky consumers who are able to get loans at 8 percent even though they would have been willing to pay 12 percent. The law is a windfall gain for them. The losers come on both the supply side and the demand side. First, there are the consumers who would have been willing and able to get credit at 12 percent but who are not lucky enough to get it at 8 percent. Then there are the banks (or, more accurately, bank stockholders) who could have made profitable loans at rates of up to 12 percent if there were no interest-rate ceiling.

This analysis helps explain the political popularity of usury laws. Few people sympathize with bank stockholders; indeed, it is the widespread feeling that banks are "gouging" their borrowers that provides much of the impetus for usury laws. The consumers who get loans at lower rates will, naturally, be quite pleased with the result of the law. The others, who would like to borrow at 8 percent but cannot because quantity supplied is less than quantity demanded,

are quite likely to blame the bank for refusing to lend, rather than blaming the government for outlawing mutually beneficial transactions.

Yet concern over high interest rates can be rational. It may, for example, be appropriate to combat homelessness by making financing of housing cheaper for poor people. But there is a difference between doing so by paying part of the necessary cost, for example, through government subsidies of interest on housing for the poor, and the alternative of declaring high costs illegal, pretending that those costs can simply be legislated away, as a usury ceiling tries to do.[4]

Whether a usury ceiling will or will not be effective depends on what the equilibrium rate of interest would have been in a free market. For example, a ceiling of 8 percent annual interest on consumer loans is quite irrelevant if the free-market equilibrium is 6 percent, but it can have important effects if the free-market rate is 12 percent.

THE DETERMINATION OF RENT: SIMPLE VERSION

The second main factor of production is land. Rent, the payment for the use of land, is another price which, when left to the market, often seems to settle at politically unpopular levels. In fact, rent controls are even more popular than usury ceilings. We discussed the effects of rent controls in Chapter 4 (pages 83–85), and we will say a bit more about them later in this chapter. But our main focus here is on the determination of rents in free markets.

The main special feature of the market for land occurs on the supply side. Land is one factor of production whose quantity supplied is (roughly) the same at every possible price. Indeed, the classical economists used this notion as the working definition of land. And the definition seems to fit, at least approximately. Although people may clear land, drain swamps, fertilize it, build on it, or convert it from one use (a farm) to another (a housing development), it is difficult to change the total supply of land very much by human effort.

What does that fact tell us about the determination of land rents? Figure 15-6 helps to provide an answer. The vertical supply curve *SS* represents the fact that no matter what the level of rents, there are still 1,000 acres of land in a small hamlet called Littletown. The demand curve *DD* is a typical marginal revenue product curve, predicated on the notion that the use of land, like everything else, is subject to diminishing returns. The free-market price is determined, as usual, by the intersection of the supply and demand curves. In this example, each acre of land in Littletown rents for $2,000 per year. The interesting feature of this diagram is that, because quantity supplied is rigidly fixed at 1,000 acres whatever the price:

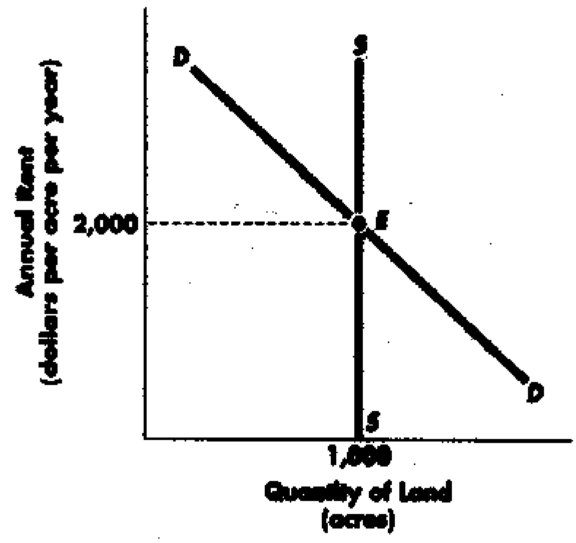
The market level of rent is entirely determined by the market's demand side.

If, for example, the relocation of a major university to Littletown attracts more people who want to live there, the *DD* curve will shift outward, as depicted in Figure 15-7. Equilibrium in the market will shift from point *E* to point *A*. There will still be only 1,000 acres of land, but now each acre will command a rent of

[4]The law also sometimes concerns itself with discrimination in lending against women or members of ethnic minority groups. There is strong evidence suggesting sex and race discrimination in lending. For example, married women have been denied loans without the explicit permission of their husbands, even where the women had substantial independent incomes.

FIGURE 15-6 DETERMINATION OF LAND RENT IN LITTLETOWN

The supply curve of land, *SS*, is vertical, meaning that 1,000 acres are available in Littletown regardless of the level of rent. The demand curve for land slopes downward for the usual reasons. Equilibrium is established at point *E*, where the annual rental rate is $2,000 per acre.



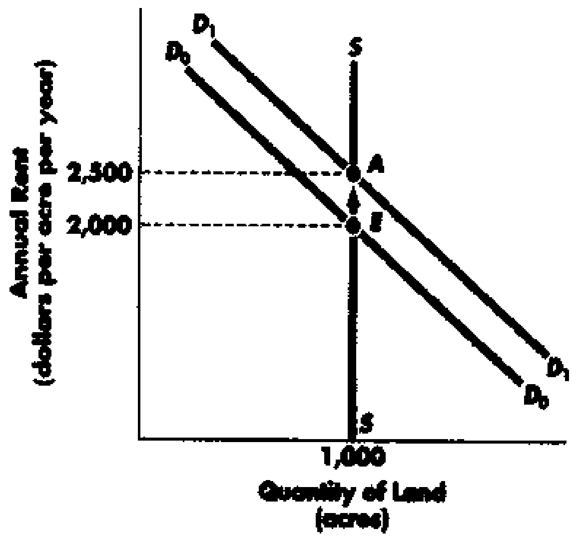
$2,500 per acre. The landlords will collect more rent, though society gets no more land from the landlords in return for its additional payment.

The same process also works in reverse, however. Should the university shut its doors and the demand for land decline as a result, the landlords will suffer even though they in no way have contributed to the decline in the demand for land. (To see this, simply reverse the logic of Figure 15-7. The demand curve begins at D_1D_1 and shifts to D_0D_0.)

This discussion shows the special feature of rent that leads economists to distinguish it from payments to other factors of production: An *economic rent* is a payment for a factor of production (such as land) that does not change the amount of that factor that is supplied—society is not compensated for a rise in its rent payments by any increase in the quantity of land it obtains.

FIGURE 15-7 A SHIFT IN DEMAND WITH A VERTICAL SUPPLY CURVE

Now imagine that something happens to increase the demand for land—that is, to shift the demand curve from D_0D_0 to D_1D_1. Quantity supplied cannot change, but the rental rate can and does. In this example, the annual rental for an acre of land increases from $2,000 to $2,500.



JAPANESE LAND PRICES

Supply and demand do not equalize prices when the commodity, like land, cannot be transferred from one geographic market to another, as the following news item illustrates:

> TOKYO (AP)—Are mortgage payments getting you down? Take heart. It could be worse. You could live in Tokyo. Tokyo residents pay 27 times as much as San Franciscans, 32 times as much as Los Angelenos and 55 times as much as some New Yorkers when they buy land for a house, according to a Japanese government survey.

SOURCE: ClariNet e.News (http://www.clari.net). Associated Press/Reuters, May 30, 1995.



THE RENT OF LAND: SOME COMPLICATIONS

If all land were of identical quality, this would be virtually all there is to the theory of land rent. But plots of land do differ—in quality of soil, in topography, in access to sun and water, in proximity to marketplaces, and in other ways. The classical economists took this into account in their analysis of rent determination—a remarkable 18th-century piece of economic logic still considered valid today.

The basic notion is that capital invested on any piece of land must yield the same return as capital invested on any other piece that is actually used. Why? If it were not so, capitalists would bid against one another for the more profitable pieces of land until the rents of these parcels were driven up to a point where their advantages over other parcels had been eliminated.

Suppose that on one piece of land a given crop is produced for $160,000 per year in labor, fertilizer, fuel, and other nonland costs, while the same crop is produced for $120,000 on a second piece of land. The rent on the second parcel must be *exactly* $40,000 per year higher than the rent on the first, because otherwise production on one plot would be cheaper than on the other. If, for example, the rent difference were only $30,000 per year, it would be $10,000 cheaper to produce on the second plot of land. No one would want to rent the first plot and every grower would instead bid for the second plot. Obviously, rent on the first plot would be forced down by lack of customers, and rent on the second would be driven up by eager bidders. These pressures would come to an end only when the rent difference reached $40,000, so that both plots became equally profitable.

At any given time, there are some pieces of land of such low quality that it does not pay to use them at all—remote deserts are a prime example. Any land that is exactly on the borderline between being used and not being used is called **marginal land.** By this definition, marginal land earns no rent because if any rent were charged for it, there would be no takers.

MARGINAL LAND is land that is just on the borderline of being used.

We now combine these two observations—that the difference between the costs of producing on any two pieces of land must equal the difference between their rents, and that zero rent is charged on marginal land—to conclude that:

Rent on any piece of land will equal the difference between the cost of producing the output on that land and the cost of producing it on marginal land.

That is, competition for the superior plots of land will permit the landlords to charge prices that capture the full advantages of their superior parcels.

A useful feature of this analysis is that it helps us to understand more completely the effects of an outward shift in the demand curve for land. Suppose population growth raises demand for land. Naturally, rents will rise. But we can be more specific than this. In response to an outward shift in the demand curve, two things will happen:

1. *It will now pay to employ some land whose use was formerly unprofitable.* The land that was previously on the zero-rent margin will no longer be on the borderline, and some land that is so poor that it was formerly not even worth considering will now just reach the borderline of profitability. The settling of the American West illustrates this process strikingly. Land that once could not be given away is now quite valuable.
2. *People will begin more intensive use of the land that was already in use.* Farmers will use more labor and fertilizer to squeeze larger amounts of crops out of their acreage, as has happened in recent decades. Urban real estate on which two-story buildings previously made most sense will now be used for high-rise buildings.

These two events will increase rents in a predictable way. Since the land that is marginal *after* the change must be inferior to the land that was marginal previously, rents must rise by the difference in yields between the old and new marginal lands. Table 15-2 illustrates this point. We deal with three pieces of land: A, a very productive piece; B, a piece that was initially marginal; and C, a piece that is inferior to B but nevertheless becomes marginal when the upward shift in the demand curve for land occurs.

The crop costs $80,000 more when produced on B than on A, and $12,000 more when produced on C than on B. Suppose, initially, that demand for the crop is so low that C is unused and B is just on the margin between being used and left idle. Since B is marginal, it will yield no rent. We know that the rent on A will be equal to the $80,000 cost advantage of A over B. Now suppose demand for the crop increases enough so that plot C is just brought into use. Plot C is now marginal land, and B acquires a rent of $12,000, the cost advantage of B over C. Plot A's rent now must rise from $80,000 to $92,000, the size of its cost advantage over C, the new marginal land.

In addition to the differences in the quality of different pieces of land, there is a second factor pushing up land rents: the increased intensity of use of land

TABLE 15-2 NONRENT COSTS AND RENT ON THREE PIECES OF LAND

Type of Land	Nonland Cost of Producing a Given Crop	Total Rent	
		Before	After
A. A tract that was better than marginal before and after	$120,000	$80,000	$92,000
B. A tract that was marginal before but is not anymore	200,000	0	12,000
C. A tract that was previously not worth using but is now marginal	212,000	0	0

that is already in cultivation. As farmers apply more fertilizer and labor to their land, the marginal productivity of land increases just as factory workers become more productive when they are given better equipment. Once again, the landowner is able to capture this increase in productivity in the form of higher rents. (If you do not understand why, refer back to Figure 15-7 and remember that the demand curves are marginal revenue product curves—that is, they indicate the amount that capitalists are willing to pay landlords for the use of their land.) Thus, we can summarize the classical theory of rent as follows:

As the use of land increases, landlords receive higher payments from two sources:

1. Increased demand leads the community to employ land previously not good enough to use; the advantage of previously used land over the new marginal land increases, and rents go up correspondingly.
2. Land is used more intensively; the marginal revenue product of land rises, thus increasing the ability of the producer who uses the land to pay rent.

As late as the end of the 19th century, this analysis still exerted a powerful influence beyond technical economic writings. An American journalist, Henry George, was nearly elected mayor of New York in 1886, running on the platform that all government should be financed by "a single tax"—a tax on landlords who, he said, are the only ones who earn incomes while contributing nothing to the productive process and who reap the fruits of economic growth without contributing to economic progress. George's logic was based on the notion that landowners do not increase the supply of their factor of production—the quantity of land—when rents increase.

GENERALIZATION: WHAT DETERMINES SHAQUILLE O'NEAL'S SALARY?

Land is not the only scarce input whose supply is fixed, at least in the short run. Toward the beginning of this century some economists realized that the economic analysis of rent can be applied to inputs other than land (see the box on the next page for some current research uses of the concept). As we will see, this extension yielded some noteworthy insights.

Consider as an example the earnings of Shaquille O'Neal, the Orlando Magic basketball star (reported to have made $30 million in 1994). Basketball players would seem to have little in common with plots of land in downtown Dallas. Yet, to an economist, the same analysis—the theory of rent—explains the incomes of both factors of production. To understand why, we first note that there is only one Shaquille O'Neal. That is, he is a scarce input whose supply is fixed just like the supply of land. Because he is in fixed supply, the price of his services is determined in a way similar to that of land rents. Hence, economists have arrived at a more general definition of **economic rent** as *any payment made to a factor above the amount necessary to induce any of that factor to be supplied to its present employment.*

ECONOMIC RENT is the portion of the earnings of a factor of production that exceeds the minimum amount necessary to induce any of that factor to be supplied.

To understand the concept of economic rent, it is, then, useful to divide the payment for any input into two parts. The first part is simply the minimum payment needed to acquire the input: the cost of producing a ball bearing or the compensation for the unpleasantness, hard work, and loss of leisure involved in performing labor. Only this first part of the factor payment is essential to induce the owner to supply any of the input. If a worker, for example, is not paid at least this first part, he or she will not supply any of this labor.

RENT SEEKING



Current research uses the rent concept to analyze such common phenomena as lobbying by industrial groups, lawsuits between rival firms, and battles over exclusive licenses (as for a television station). Such interfirm battles can waste very valuable economic resources, for example, the time spent by executives, bureaucrats, judges, lawyers, and economists. Because this valuable time could have been used in production, such activities entail large *opportunity costs*. The new analysis offers insights into the reasons for these battles and provides a way to assess what *quantity* of resources is wasted.

What is the relevance of economic rent—a payment to a factor of production above and beyond the amount necessary to get the factor to make its contribution to production? The search and battle for such opportunities is called "rent seeking," a concept introduced by Gordon Tullock, an economist trained in legal matters.

An obvious source of such rents is a monopoly license. For example, a license to operate the only TV station in town will yield enormous advertising profits, far above the amount needed to induce the station to operate. No wonder rent seekers swoop down when such a license becomes available. Similarly, the powerful lobby of U.S. producers of sweeteners pressures Congress to impede imports of cane sugar, since free importation would cut prices (and rents) substantially.

How much of society's resources will be wasted in such a process? The theory of rent seeking gives us some idea. Thus, consider a race for a monopoly cable TV license which, once awarded, will keep competitors out. But nothing prevents anyone from entering the race to *grab* the license. Anyone can hire the lobbyists and lawyers or offer the bribes needed in the battle. Thus, while the cable business is itself not competitive, the process of fighting for the license is.

But, we know from the analysis of long-run equilibrium under perfect competition (Chapter 9, pages 221–225) that economic profit approximates zero—in other words, revenues just cover costs. So, if the cable license is expected to yield, over its life, say, $900 million in rent, rent seekers are likely to waste something near that amount in the fight for the license.

Why? Suppose there are ten bidders, each with an equal chance at the prize. Then, to each bidder that chance should be worth about $90 million. If the average bidder has so far spent, say, only $70 million on the battle, there will still be an expected economic profit of $90 million minus $70 million, or $20 million to the rent-seeking activity. This will tempt an 11th bidder to enter and raise the ante, say, to $80 million in lobbying fees, hoping to grab the rent. This process stops only when enough of the rent has been wasted on the rent-seeking process.

The second part of the payment is a bonus that does not go to every input, but only to those that are of particularly high quality. Payments to workers with exceptional natural skills are a good example. These bonuses are like the extra payment for a better piece of land, and so are called *economic rents*. Indeed, like the rent of land, an increase in the amount of economic rent paid to an input may not increase the quantity of that input supplied. This second part—the economic rent—is pure gravy. The skillful worker is happy to have it as an extra. But it is not a deciding consideration in the choice of whether or not to work.

A moment's thought shows how this general notion of rent applies both to land and to Shaquille O'Neal. The total quantity of land available for use is the same whether rent is high, low, or zero; no payments to landlords are necessary to induce land to be supplied to the market. So, by definition, the payments to landholders for their land are entirely economic rent—payments that are not necessary to induce the provision of the land to the economy. Shaquille O'Neal is (almost) similar to land in this respect. His athletic talents are somewhat unique and cannot be reproduced. What determines the income of such a factor? Since the quantity supplied of such a unique, nonreproducible factor is absolutely fixed, and therefore unresponsive to price, the analysis of rent determination summarized in Figure 15-6 applies. *The position of the demand curve determines the price.*

Figure 15-8 summarizes the "Shaquille O'Neal market." The mostly vertical supply curve *TUR* represents the fact that no matter what wage he is paid, there is only one Shaquille O'Neal. Demand curve *DD* is a marginal productivity curve of sorts, but not quite the kind we encountered earlier in the chapter. Since the question, "What would be the value of a second unit of Shaquille O'Neal?" is nonsensical, we construct the demand curve by considering only the *portion* of his time demanded at various wage levels. The curve indicates that at an annual salary of $50 million no employer can afford even a little bit of O'Neal. At a lower salary of, say, $35 million per year, however, two-thirds of his time will be demanded. At $30 million per year, O'Neal's full time is demanded; and at lower wage rates, like $25 million, the demand for his time exceeds the amount of it that is for sale.

Equilibrium is at point *E* in the diagram, where the supply of and demand for his time are equal. His annual salary here is $30 million. Now we can ask: How much of Shaquille O'Neal's salary is economic rent? According to the economic definition of rent, only part of his $30 million salary is rent. Because the supply schedule is only partly vertical, part of O'Neal's financial reward is necessary to get him to supply his services, as is undoubtedly true in reality. Thus, it is not true that every penny he earns is rent.

This is why we said that top athletes like Shaquille O'Neal are *almost* good examples of pure rent. For, in fact, if his salary were low enough, O'Neal might well prefer to play golf rather than work. Suppose, for example, that $250,000 per year is the lowest salary at which O'Neal will offer even 1 minute of his services, and that his labor supply then increases with his wage up to an annual

FIGURE 15-8 HYPOTHETICAL MARKET FOR SHAQUILLE O'NEAL'S SERVICES

At an annual wage of $50 million or more, no one is willing to bid for O'Neal's time. At a lower wage, $35 million, two-thirds of his time will be demanded (point *G*). Only at an annual wage no higher than $30 million will all of O'Neal's available time be demanded (point *E*).

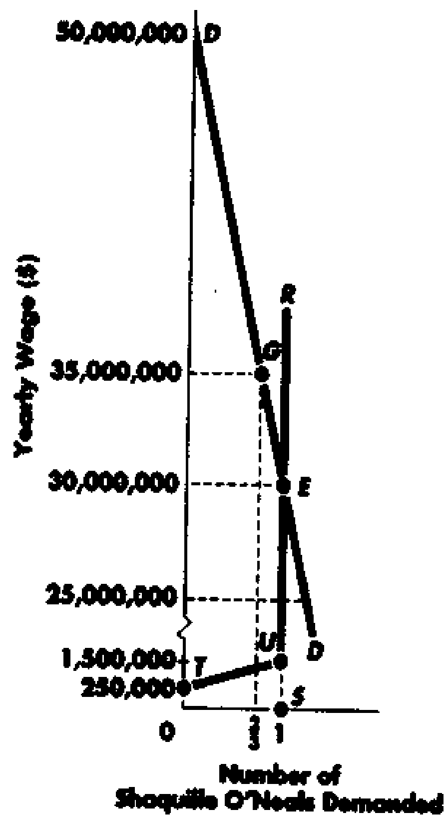
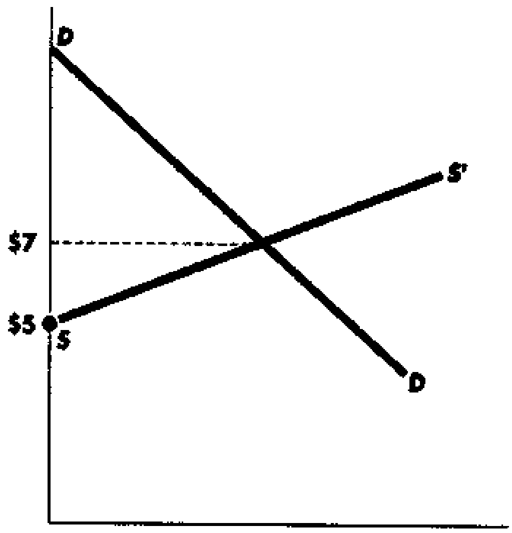


FIGURE 15-9 RENT WHEN THE SUPPLY CURVE IS NOT VERTICAL

Some of the input would be supplied (point *S*) at a price of $5 (or a bit more), but the equilibrium price is $7 so some units of input must be earning a rent of $2.



salary of $1,500,000, at which point he is willing to work full time. Then, while his equilibrium salary will still be $30 million per year, not all of it will be rent, because some of it, at least $250,000, is required to get him to supply any services at all.

The portion of Shaquille O'Neal's compensation that is not pure rent corresponds to the upward-sloping portion, *TU*, of his labor-supply curve. In Figure 15-8 that is why only part of the supply curve, *TUR*, is vertical—that portion above the $1,500,000 salary that will lead him to supply all his available time. And his equilibrium compensation level, *E*, will consist partly of rent (portion *UE*) and partly of a payment, *SU*, that is not rent.

This same analysis applies to any factor of production whose supply curve is not horizontal, as in Figure 15-9. There we see that at any price above $5, suppliers are willing to provide some units of the input; that is, at any price above point *S*, quantity supplied is greater than zero. Yet the supply-demand equilibrium point yields a price of $7—well above the minimum price at which some input supply would be forthcoming. The difference must constitute a rent to the input suppliers, who get paid more than the minimum amount required to induce them to work.

Almost all employees earn some rent. What sorts of factors earn no rent? Those that can be exactly reproduced by a number of producers at constant cost. No supplier of ball bearings will ever receive any rent on a ball bearing, at least in the long run, because any desired number of them can be produced at (roughly) constant costs—say, 5 cents each. If one supplier tried to charge a price above 5 cents, another manufacturer would undercut the first supplier and take its customers away. Hence, the competitive price will include no rent.

RENT CONTROLS: THE MISPLACED ANALOGY

Why is the analysis of economic rent important? Because only economic rent can be taxed away without reducing the quantity of the input supplied. And here common English gets in the way of sound reasoning. Many people feel that the

rent that they pay to their landlord is economic rent. After all, their apartments will still be there if they pay $500 per month, or $300, or $100. This view, while true in the short run, is quite myopic.

Like the ball-bearing producer, the owner of a building cannot expect to earn *economic* rent because there are too many other potential owners whose costs of construction are roughly the same as her own. If the market price temporarily included some economic rent—that is, if price exceeded production costs plus the opportunity cost of the required capital—other builders would start new construction that would drive the price down. Thus, far from being in perfectly *inelastic* (vertical) supply, like raw land, buildings come rather close to being in perfectly *elastic* (horizontal) supply, like ball bearings. As we have learned from the theory of rent, this means that builders and owners of buildings cannot collect economic rent in the long run.

Since apartment owners collect very little economic rent, payments by tenants in a free market must be just enough to keep those apartments on the market. (This is the definition of zero economic rent.) If rent controls push these prices down, the apartments will start disappearing from the market.[5] Among other unfortunate results, we can therefore expect rent controls to contribute to homelessness—though it is, of course, not the only influence behind this distressing phenomenon.

ISSUE: ARE PROFITS TOO HIGH OR TOO LOW?

This completes our analysis of rent. We turn next to business profits, a subject whose discussion seems to elicit more passion than logic. With the exception of some economists, almost no one thinks that the rate of profit is at about the right level. Critics point accusingly at the billion-dollar profits of some giant corporations and argue that they are unconscionably high. They call for much stiffer profits taxes. On the other hand, the Chambers of Commerce, National Association of Manufacturers, and other business groups complain that regulations and "ruinous" competition keep profits too low, and they are constantly petitioning Congress for tax relief.

The public has many misconceptions about the nature of the U.S. economy, but probably none is more severe than the popular view of the amount of profit that American corporations earn. Try the following experiment. Ask five of your friends who have never had an economics course what fraction of the nation's income they imagine is accounted for by profits. While the correct answer varies from year to year, in 1995 about 8.4 percent of GDP (before-tax) was business profits.[6] A comparable percentage of the prices you pay represents before-tax profit. Most people think this figure is much, much higher—as we saw in the box in Chapter 2 on page 35.

As you have no doubt noticed by now, economists are reluctant to brand factor prices as "too low" or "too high" in some moral or ethical sense. Rather, they

[5]None of this is meant to imply that temporary rent controls in certain locations cannot have desirable effects in the short run. In the short run, the supply of apartments and houses really is fixed, and large shifts in demand can hand windfall gains to landlords—gains that are true economic rents. Controls that eliminate such windfalls should not cause serious problems. But knowing when the "short run" fades into the "long run" can be tricky. "Temporary" rent control laws have a way of becoming rather permanent.

[6]*Economic Report of the President 1996* (Washington, D.C.: U.S. Government Printing Office, February 1996), pp. 280 and 379.

are likely to ask first: What is the market equilibrium price? And then they will ask whether there are any good reasons to interfere with the market solution. This analysis, however, is not so easy to apply to the case of profits, since it is hard to use supply and demand analysis when you do not know what factor of production earns profit.

In both a bookkeeping and an economic sense, *profits are the residual*. They are what remains from the selling price after all other factors have been paid.

But what factor of production receives this reward? What factor's marginal productivity constitutes the profit rate?

WHAT ACCOUNTS FOR PROFITS?

Economic profit, as we learned in Chapter 9, is the amount a firm earns *over and above* the payments for all other inputs, including the interest payments for the capital it uses and the opportunity cost of any capital provided by the owners of the firm. The payment of the opportunity cost of the capital that the owners provide to the firm (that in common parlance is considered profit) is closely related to the interest rate. In an imaginary (and uninteresting) world in which everything was certain and unchanging, capitalists who invested money in firms would simply earn the market rate of interest on their funds. Profits beyond this level would be competed away. Payment for capital below this level could not persist, because capitalists would withdraw their funds from firms and deposit them in banks. Capitalists in such a world would be mere moneylenders.

But the real world is not at all like this. Some capitalists are much more than moneylenders, and the amounts they earn often exceed the interest rate by a considerable margin. These activist capitalists who seek out or even create earnings opportunities are called *entrepreneurs*. They are the ones who are responsible for the constant change that characterizes business firms and who prevent the operations of the firms from stagnating (see the box, opposite). Since they are always trying to do something new, it is difficult to provide a general description of their activities. However, we can list three primary ways in which entrepreneurs are able to drive profits above the level of interest rates.

EXERCISE OF MONOPOLY POWER

If the entrepreneur can establish a monopoly over some or all of his products, even for a short while, he can use the monopoly power of his firm to earn monopoly profits. We analyzed the nature of these monopoly earnings in Chapter 11.

RISK BEARING

The entrepreneur may engage in risky activities. For example, when a firm prospects for oil, it will drill an exploratory well hoping to find a pool of petroleum at the bottom. But a high proportion of such attempts produces only dry holes, and the cost of such an operation is wasted. Of course, if the investor is lucky and does find oil, she may be rewarded handsomely. The income she obtains is a payment for bearing risk.

Obviously, a few lucky individuals make out well in this process, while most suffer heavy losses. How well can we expect risk takers to do on the average? If, on the average, one exploratory drilling out of ten pays off, do we expect its return to be exactly 10 times as high as the interest rate, so that the *average* firm

ENTREPRENEURSHIP IN POLAND

Entrepreneurship is not easy to find or create, as the following story illustrates:

> Maybe it was beginnner's luck, but a lot of would-be capitalists got off to a fast start after the fall of the Berlin Wall in 1989. Qumak International, then a fledgling computer importer, was one of them. It couldn't fill orders fast enough. Revenue growth was [illegible]; the company grossed $2 million in 1990 and $8 million in 1994. But Polish entrepreneurs are now in the process of learning about another fundamental aspect of western-style capitalism: Big bottom-line profits attract a lot of competition.
>
> Qumak, and a lot of other new companies in eastern Europe, now face a host of new rivals. Revenues fell about 25 percent and the company is strapped by cash-flow problems. One-fifth of the work force has been let go. Foreign competitors are one culprit, and high interest rates have added to the difficulty.
>
> All of these pitfalls combine to indicate other fundamental requirements of free-market economies. Business enterprises must be skillfully managed in order to survive over the long haul. And entrepreneurs, even those with hot new products or groundbreaking services, can expect plenty of competitive imitators in short order.
>
> Qumak has already been forced to revamp its initial strategy. It has abandoned the crowded personal computer market and instead focused on the more lucrative business computer sector. If the company is to grow from a small operation into a large, stable company, many more such managerial decisions are in store. Entrepreneurs are discovering that fast starts are nice, but surviving in a free-market economy requires careful management. It also pays to keep a close eye on the competition.



SOURCE: Dana Milbank, "Tough Business: Polish Entrepreneurs Revitalize Economy but Battle Huge Odds," *The Wall Street Journal*, March 30, 1995, pp. A1 and A6.

will earn exactly the normal rate of interest? The answer is that the payoff will be *more* than 10 times the interest rate if investors dislike gambling, that is, if they prefer to avoid risk. Why? Because investors who dislike risk will be unwilling to put their money into a business in which nine firms out of ten lose out unless there is some compensation for the financial peril.

In reality, however, there is no certainty that things always work out this way. Some people love to gamble and tend to be overoptimistic. They may plunge into projects to a degree unjustified by the odds. If there are enough such gamblers, the average payoff to risky undertakings may end up below the interest rate. The successful investor will still make a good profit, just like the lucky winner in Las Vegas. But the average participant will have to pay for the privilege of bearing risk.

RETURNS TO INNOVATION

INVENTION is the act of generating an idea for a new product or a new method for making an old product.

INNOVATION, the next step, is the act of putting the new idea into practical use.

The third major source of profits is perhaps the most important of all from the point of view of social welfare. The entrepreneur who is first to market a desirable new product or employ a new cost-saving machine will receive a profit higher than that normally accruing to an uninnovative (but otherwise similar) business manager. Innovation is different from invention. **Invention** is the act of generating a new idea; **innovation** is the next step, the act of putting the new idea into practical use. Business people are rarely inventors, but they are often innovators.

When an entrepreneur innovates, even if her new product or new process is not protected by patents, she will be one step ahead of her competitors. She will

be able to capture much of the market either by offering customers a better product or by supplying the product more cheaply. In either case she will temporarily find herself with some monopoly power left by the weakening of her competitors, and monopoly profit will be the reward for her initiative.

However, this monopoly profit, the reward for innovation, will only be temporary. As soon as the success of the idea has demonstrated itself to the world, other firms will find ways of imitating it. Even if they cannot turn out precisely the same product or the same process, they must find close substitutes in order to survive. In this way, new ideas spread through the economy. And in the process the special profits of the innovator are brought to an end. The innovator can only resume earning special profits by finding still another promising idea.

Entrepreneurs are forced to keep searching for new ideas, to keep instituting innovations, and to keep imitating those ideas that they were not the first to put into operation. This process is at the heart of the growth of the capitalist system. It is one of the secrets of its extraordinary dynamism.

THE ISSUE OF PROFITS TAXATION

So profits in excess of the market rate of interest can be considered as the return on entrepreneurial talent. But this is not really very helpful, since no one can say exactly what entrepreneurial talent is. Certainly we cannot measure it; nor can we teach it in a college course. (Business schools try, though!) Therefore, we do not know how the observed profit rate relates to the minimum reward necessary to attract entrepreneurial talent into the market—a relationship that is crucial for the contentious issue of profits taxation.

Consider the windfall profits tax on oil companies as an example. If oil company profit rates are well above this minimum, they contain a large element of economic rent. In that case, we could tax away these excess profits (rents) without fear of reducing oil production. On the other hand, if the profits being earned by oil companies do not contain much economic rent, then the windfall profits tax can seriously curtail exploration and production of oil.

This example illustrates the general problem of deciding how heavily profits should be taxed. Critics of big business who call for high, if not confiscatory, profits taxes seem to believe that profits are mostly economic rent. But if they are wrong, if most of the observed profits are necessary to attract people into entrepreneurial roles, then a high profits tax can be dangerous. It can threaten the very lifeblood of the capitalist system. Business lobbying groups predictably claim that this is the case. Unfortunately, neither group has offered much evidence for its conclusion.

CRITICISMS OF MARGINAL PRODUCTIVITY THEORY

The theory of factor pricing described in this chapter is another example of supply-demand analysis. Its special feature is its heavy reliance on the principle of marginal productivity to derive the shape and position of the demand curve. For this reason, the analysis is often rather misleadingly called the *marginal productivity theory of distribution*, when it is, at best, only a theory of the demand side of the pertinent market.

Over the years, this analysis has been subject to attack on many grounds. One frequent accusation, which is largely (but not entirely) groundless, is the assertion that marginal productivity theory is merely an attempt to justify the distri-

bution of income that the capitalist system yields—that it is a piece of pro-capitalist propaganda. According to this argument, when marginal productivity theory claims that each factor is paid exactly its marginal revenue product, this is only a sneaky way of asserting that each factor is paid exactly what it deserves. These critics claim that the theory legitimizes the gross inequities of the system—the poverty of many and the great wealth of the few.

The argument is straightforward but wrong. Payments are made not to *factors of production* but to the people who happen to own them. If an acre of land earns $2,000 because that is its marginal revenue product, this does not mean, nor is it meant to imply, that the payment is *deserved* by the landlord, who may even have acquired the land by fraud.

Second, an input's marginal revenue product (MRP) does not depend only on "how hard it works" but also on how much of it happens to be employed—for, according to the "law" of diminishing returns, the more that is employed, the lower its MRP. Thus, that factor's MRP is not and cannot legitimately be interpreted as a measure of the intensity of its "productive effort." In any event, what an input deserves, in some moral sense, may depend on more than what it does in the factory. For example, workers who are sick or have many children may be more "deserving," even if they are no more productive.

On these and other grounds, no economist today claims that marginal productivity analysis shows that distribution under capitalism is either just or unjust. It is simply wrong to claim that marginal productivity theory is pro-capitalist propaganda. The marginal productivity principle is just as relevant to organizing production in a socialist society as it is in a capitalist one.

Others have attacked marginal productivity theory for using rather complicated reasoning to tell us very little about the really urgent problems of income distribution. In this view, it is all very well to say that everything depends on supply and demand and to express this in terms of many complicated equations (as is done in more advanced books and articles). But these equations do not tell us what to do about such serious distribution problems as malnutrition among the indigenous populations in Latin America or poverty among minority groups in the United States.

Though it does exaggerate somewhat, there is certainly truth to this criticism. We have seen in this chapter that the theory does provide some insights on real policy matters, though not as many as we would like. Later in the book we will see that economists do have things to say about the problems of poverty and underdevelopment. But much of this does not flow from marginal productivity analysis.

Perhaps, in the end, what should be said for marginal productivity theory is that it is the best model we have at the moment, that it offers us *some* valuable insights into the way the economy works, and that until a more powerful model is found we are better off hanging on to what we have.

SUMMARY

1. A profit-maximizing firm purchases that quantity of any input at which the price of the input equals its **marginal revenue product**. Consequently, the firm's demand curve for an input is the downward-sloping portion of that input's curve.
2. Interest rates are determined by the supply of and demand for funds. The demand for funds is a derived demand, since these funds are used to finance business **investment**. Thus, the demand for funds depends on the marginal productivity of **capital**.

3. A dollar obtainable sooner is worth more than a dollar obtainable later because of the **interest** that can be earned in the interim.
4. Increased demand for a good that needs land to produce it will drive up the prices of land either because inferior land will be brought into use or because land will be used more intensively.
5. Rent controls do not significantly affect the supply of land, but they do tend to reduce the supply of buildings.
6. **Economic rent** is any payment to the supplier of a factor of production that is greater than the minimum amount needed to induce any of the factor to be supplied.
7. Factors of production that are unique in quality and difficult or impossible to reproduce will tend to be paid relatively high **economic rents** because of their scarcity.
8. Factors of production that are easy to produce at a constant cost and that are provided by many suppliers will earn little or no economic rent.
9. Economic profits over and above the cost of capital are earned (a) by exercise of monopoly power, (b) as payments for bearing risk, and (c) as the earnings of successful **innovation.**
10. The desirability of increased taxation of profits depends on its effects on the supply of entrepreneurial talent. If most profits are economic rents, then higher profits taxes will have few detrimental effects. But if most profits are necessary to attract **entrepreneurs** into the market, then higher profits taxes can weaken the capitalist system.

KEY TERMS

Factors of production
Entrepreneurship
Marginal productivity principle
Marginal physical product (MPP)
Marginal revenue product (MRP)
Derived demand
Investment
Capital
Interest
Discounting
Usury law
Marginal land
Economic rent
Entrepreneurs
Risk bearing
Invention versus innovation

QUESTIONS FOR REVIEW

1. A profit-maximizing firm expands its purchase of any input up to the point where diminishing returns have reduced the marginal revenue product so that it equals the input price. Why does it not pay the firm to "quit while it is ahead," buying so small a quantity of the input that the input's MRP remains greater than its price?
2. Which of the following inputs do you think include relatively large economic rents in their earnings?
   a. Nuts and bolts
   b. Petroleum
   c. A champion racehorse

   Use supply-demand analysis to explain your answer.
3. Three machines are employed in an isolated area. They each produce 2,000 units of output per month, the first requiring $20,000 in raw materials, the second $25,000, and the third $28,000. What would you expect to be the monthly charge for the first and second machines if the services of the third machine can be hired at a price of $9,000 a month? What parts of the charges for the first two machines are economic rent?
4. Economists conclude that a tax on the revenues of firms will be shifted in part to consumers of the products of those firms in the form of higher product prices. However, they believe that a tax on the rent of land usually cannot be shifted. What explains the difference?
5. Many economists argue that a tax on apartment houses is likely to reduce the supply of apartments, but that a tax on all land, including the land on which apartment houses stand, will not reduce the supply of apartments. Can you explain the difference? How is this answer related to the answer to Review Question 4?
6. Distinguish between investment and capital.
7. If you have a contract under which you will be paid $10,000 in 2 years from now, why do you become richer if the rate of interest falls?
8. What is the difference between interest and profit? Who earns interest, in return for what contribution to production? Who earns economic profit, in return for what contribution to production?
9. Do you know any entrepreneurs? How do they earn a living? How do they differ from managers?
10. Explain the difference between an invention and an innovation. Give an example of each.

11. "Marginal productivity does not determine how much a worker will earn—it determines only how many workers will be hired at a given wage. Therefore, marginal productivity analysis is a theory of demand for labor, not a theory of distribution." What, then, do you think determines wages? Does marginal productivity affect their level? If so, how?
12. (More difficult) American savings rates are among the lowest of any industrial country. This has caused concern about our ability to finance new plants and equipment for U.S. industry. Some politicians and others have advocated lower taxes on saving as a remedy. Do you expect such a program to be very effective? Why?
13. If rent constitutes less than 1 percent of the incomes of Americans, why may the concept nevertheless be significant?
14. Litigation in which one company sues another often involves costs for lawyers and other court costs literally amounting to hundreds of millions of dollars per case. What does rent have to do with the matter?
15. (More difficult) This chapter (pages 358–359) explained that a rise in interest rates will increase some people's saving (as one might expect), but will decrease other people's saving because it reduces the amount of saving they need to reach some target. Analyze the consequences of an increase in the interest rate in terms of its income and substitution effects. (For review of these concepts see Chapter 5, pages 108–110.)

APPENDIX DISCOUNTING AND PRESENT VALUE

Frequently, in business and economic problems, it is necessary to compare sums of money received (or paid) at different dates. Consider, for example, the purchase of a machine that costs $11,000 and will yield a marginal revenue product of $14,520 in 2 years from today. If the machine can be financed by a 2-year loan bearing 10 percent interest, it will cost the firm $1,100 in interest at the end of each year, plus $11,000 in principal repayment at the end of the second year. (See the table that follows.) Is the machine a good investment?

COSTS AND BENEFITS OF INVESTING IN A MACHINE

	End of Year 1	End of Year 2
Benefits		
Marginal revenue product of the machine	$ 0	$14,520
Costs		
Interest	1,100	1,100
Repayment of principal on loan	0	11,000
Total	1,100	12,100

The total costs of owning the machine over the 2-year period ($1,100 + $12,100 = $13,200) are less than the total benefits ($14,520). But this is clearly an invalid comparison, because the $14,520 in future benefits are not worth $14,520 in terms of today's money. Adding up dollars received (or paid) at different dates is a bit like adding apples and oranges. The process that has been invented for making these magnitudes comparable is called *discounting*, or *computing the present value* of a future sum of money.

To illustrate the concept of present value, let us ask how much $1 received a year from today is worth *in terms of today's money*. If the rate of interest is 10 percent, the answer is about 91 cents. Why? Because if we invest 91 cents today at 10 percent interest, it will grow to 91 cents plus 9.1 cents in interest = 100.1 cents in a year. Similar considerations apply to any rate of interest. In general:

If the rate of interest is i, the present value of $1 to be received in a year is:

$$\frac{\$1.00}{(1 + i)}$$

This is so, because in a year $\frac{\$1.00}{(1+i)}$ will grow to:

$$\frac{\$1.00}{1+i} \times (1 + i) = \$1$$

What about money to be received 2 years from today? Using the same reasoning, and supposing the interest rate is 10 percent so that $1 + i = 1.1$, $1.00 invested today will grow to $1.00 times (1.1) = $1.10 after 1 year and to $1.00 times (1.1) times (1.1) = $1.00 times $(1.1)^2$ = $1.21 after 2 years. Consequently, the present value of $1.00 to be received 2 years from today is:

$$\frac{\$1.00}{(1+i)^2} = \frac{\$1.00}{1.21} = 82.64 \text{ cents}$$

A similar analysis applies to money received 3 years from today, 4 years from today, and so on.

The general formula for the present value of $1.00 to be received N years from today when the rate of interest is i is:

$$\frac{\$1.00}{(1+i)^N}$$

The present value formula highlights the two variables that determine the present value of any future flow of money: the rate of interest (i) and how long you have to wait before you get it (N).

Let us now apply this analysis to our example. The present value of the revenue is easy to calculate since it all comes 2 years from today. Since the rate of interest is assumed to be 10 percent ($i = 0.1$) we have:

$$\text{Present value of revenues} = \frac{\$14{,}520}{(1.1)^2} = \frac{\$14{,}520}{1.21} = \$12{,}000$$

The present value of the costs is a bit trickier in this example since costs occur at two different dates.

The present value of the first interest payment is $\frac{\$1{,}100}{(1+i)} = \frac{\$1{,}100}{1.1} = \$1{,}000$. And the present value of the final payment of interest plus principal is:

$$\frac{\$12{,}100}{(1+i)^2} = \frac{\$12{,}100}{(1.1)^2} = \frac{\$12{,}100}{1.21} = \$10{,}000$$

Now that we have expressed each sum in terms of its present value, it is permissible to add them up. So the present value of all costs is:

$$\text{Present value of costs} = \$1{,}000 + \$10{,}000 = \$11{,}000$$

Comparing this figure to the $12,000 present value of the revenues clearly shows that the machine really is a good investment. We can use the same calculation procedure for all investment decisions.

SUMMARY

1. To determine whether a loss or a gain will result from a decision whose costs and returns will come at several different periods of time, we must discount all the figures represented by these gains and losses to obtain their present value.
2. For this, we use the present value formula for X dollars receivable N years from now:

$$\text{Present value} = \frac{X}{(1+i)^N}$$

3. We then add together the present values of all the returns and all the costs. If the sum of the present values of the returns is greater than the sum of the present values of the costs, then the decision to invest will promise a net gain.

KEY TERMS

Discounting
Present value

QUESTIONS FOR REVIEW

1. Compute the present value of $1,000 to be received in 3 years if the rate of interest is 11 percent.
2. A government bond pays $100 in interest each year for 3 years and also returns the principal of $1,000 in the third year. How much is it worth in terms of today's money if the rate of interest is 8 percent? If the rate of interest is 12 percent?

CHAPTER 16

LABOR: THE HUMAN INPUT

Octavius (a wealthy young Englishman): "I believe most intensely in the dignity of labor."
The chauffeur: "That's because you never done any."
George Bernard Shaw, *Man and Superman*, Act II

Labor costs account, by far, for the largest share of GDP. As noted in the previous chapter, the earnings of labor amount to almost 75 percent of national income. Wages also represent the primary source of income to the vast majority of Americans.

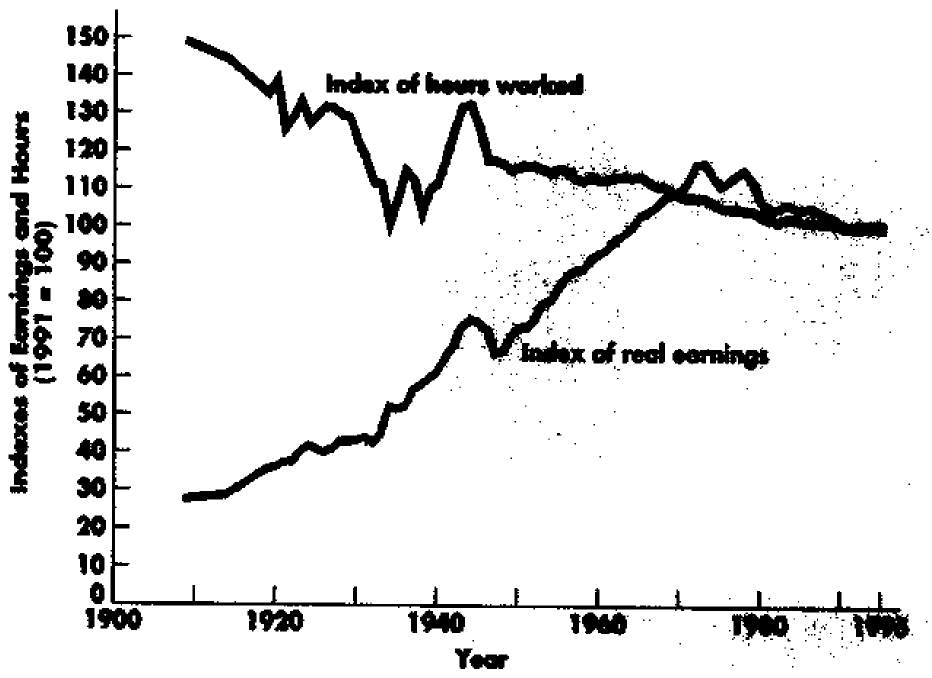
The first part of this chapter deals with the determination of wages and employment in *competitive labor markets*, that is, labor markets in which there are many buyers and many sellers, none of whom is large enough to have any appreciable influence on wages. We consider why some types of workers are paid far more than others. In the second part of the chapter we discuss labor markets that are monopolized on the selling side by trade unions. Finally, we turn to situations in which a single seller of labor (a union) confronts a single buyer of labor (a monopsony firm), and examine some of the analytical and practical difficulties that arise under collective bargaining.

ISSUE: LAGGING WAGES AND GROWING INEQUALITY

For about two decades, real (inflation-adjusted) wages and, with them, the distribution of income, have shown disturbing trends that differ sharply from America's past. As Figure 16-1 shows, for the first three-quarters of the 20th century real hourly wages rose sharply and quite steadily. This upward march was a major foundation of the American dream. It offered the prospect that members of the labor force could expect tomorrow's earnings to buy more and better goods and services than today's, and that their children and grandchildren could be expected to do better still.

Then suddenly, beginning around 1973, real wages stopped rising. Adding in employer payments for fringe benefits such as pensions and health insurance brightens the picture somewhat, but not very much. Wages and fringe benefits together just about remained level, and they certainly did not continue the remarkable growth trend of earlier decades. Besides, rising fringe benefits were in part a response to sharply increasing costs of services such as health care rather than an increase in the quantity and quality of the services provided to workers.

FIGURE 16-1 TRENDS IN REAL WAGES AND HOURS WORKED, 1909–1995



This graph shows how real wages in the United States (measured in 1982–1984 purchasing power dollars) increased during the first seven decades of the 20th century and then stopped rising, while hours worked per week declined, despite the higher rewards for each hour of work. The sharp drop in hours during the 1930s reflects the high unemployment of the Great Depression, and the sharp rise in hours in the 1940s reflects the unusual circumstances of World War II.

SOURCES: Constructed by the authors from data in Bureau of the Census, *Historical Statistics of the United States, Colonial Times to 1970* (Washington, D.C.: U.S. Government Printing Office, 1975); and *Economic Report of the President* (Washington, D.C.: U.S. Government Printing Office, various years). Data on both weekly hours and hourly wages pertain to the entire civilian economy for the period since 1947, but only to the manufacturing sector for earlier years because of the unavailability of economywide data.

Slowing wage growth was accompanied by an expanding gap between rich and poor. Today, high-wage families in the United States reportedly earn four times as much as low-wage families—and the disparity is rising. In Germany, by contrast, the ratio between low-wage and high-wage families is only 2.5. And, as one economist recently pointed out, "We are the most unequal industrialized country in terms of income and wealth, and we're growing unequal faster than the other industrialized countries."[1] With 20 percent of its children living below the poverty line, the United States has more than double the child poverty rate of any other industrialized country.[2]

These developments have profound, and mostly distressing, implications for the future of our society. We are not sure what is behind these trends; nor do we quite know what can be done about them. But we will discuss some of the potential causes later in the chapter.

COMPETITIVE LABOR MARKETS

In a completely free labor market, wages (the price of labor) would be determined by supply and demand, just like any other price. But the labor market has a number of distinctive features, mostly on the supply side. On the demand side, we will find that the demand curve for labor is derived like the demand curve for any other input—by labor's marginal revenue product.

[1] "Studies Show Major Economic Division," Reuters, April 17, 1995 (from *The New York Times*). For more information on this issue, see Edward N. Wolff, *Top Heavy: A Study of Increasing Inequality of Wealth in America* (New York: Twentieth Century Fund, 1995).

[2] United Nations International Children's Emergency Fund (UNICEF), *The Progress of Nations* (New York: United Nations, 1993), p. 45.

The labor market is also generally far from perfectly competitive. Nonetheless, we start our investigation by describing the theory of competitive labor markets in which the buyers are large numbers of tiny firms and the sellers are individual workers. Both buyers and sellers are too small to have any choice but to accept the wage rate determined by the impersonal forces of supply and demand.

THE SUPPLY OF LABOR

The economic analysis of labor supply makes extensive use of a simple observation: Given the fixed amount of time in a week, a person's decision to *supply labor* to firms is simultaneously a decision to *demand leisure* time for herself. Assuming that, after deducting the necessary time for eating and sleeping, a worker has 90 usable hours in a week, then a decision to spend 40 of those hours working is simultaneously a decision to demand 50 of them for other purposes.

This suggests that we can analyze the *supply* of labor with the same tools we used in Chapter 5 to analyze the *demand* for commodities. In this case, the commodity is leisure. Consumers "buy" their own leisure time, just as they buy bananas or back scratchers or bicycles. In Chapter 5 we observed that any price change has two distinct effects on quantity demanded: an income effect and a substitution effect. Let us review these two effects and see how they affect the demand for leisure (that is, the supply of labor). We will see that they tell us a good deal about the labor market, as well as the markets for some other inputs.[3]

1. *Substitution effect.* Consumers "purchase" their own leisure time by giving up their hourly wage, so the wage rate is the "price" (the opportunity cost) of leisure. When the wage rate rises, leisure becomes more expensive relative to other commodities that consumers might buy. This leads us to expect a wage increase to induce them to buy *less* leisure time and *more* goods. Thus:

   The substitution effect of higher wages probably leads most workers to want to work more.

2. *Income effect.* Higher wages make consumers richer. We expect this increased wealth to raise the demand for most goods, *leisure included.* So:

   The income effect of higher wages probably leads most workers to want to work less, while lower wages make them want to work more.

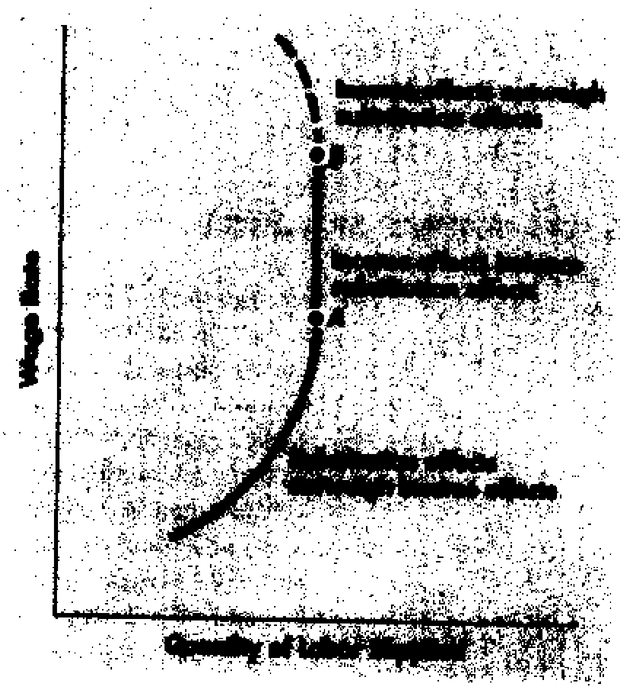
Putting these two effects together, we are led to conclude that some workers may react to an increase in their wage rate by working more, while others may react by working less. Still others will have little or no discretion over their hours of work. In terms of the market as a whole, therefore, higher wages can lead to either a larger or a smaller quantity of labor supplied.

Statistical studies of this issue in the United States have reached the conclusions that (a) the response of labor supply to wage changes is not very strong for most workers; (b) for low-wage workers the substitution effect seems clearly dominant, so they work more when wages rise; and (c) for high-wage workers the income effect just about offsets the substitution effect, so they do not work more when wages rise. Figure 16-2 depicts these approximate "facts." It shows

[3]See Review Question 15 at the end of the previous chapter.

FIGURE 16-2 A TYPICAL LABOR SUPPLY SCHEDULE

The labor supply schedule depicted here has a positive slope up to point *A*, as substitution effects outweigh income effects. At higher wages, however, income effects become just as important as substitution effects, and the curve becomes roughly vertical. At still higher wages (above point *B*), income effects might overwhelm substitution effects.



labor supply rising (slightly) as wages rise up to point *A*. Thereafter, labor supply is roughly constant as wages rise.

It is even possible that when wages are raised high enough, further wage increases will lead workers to purchase more leisure and therefore to work less. The supply curve of labor is then said to be a *backward-bending supply curve*, as illustrated by the broken portion of the curve above point *B*.

Does the theory of labor supply apply to college students? A study of the hours worked by students at Princeton University found that it does.[4] Estimated substitution effects of higher wages on the labor supply of Princeton University students were positive and income effects were negative, just as the theory predicts. Apparently, substitution effects outweighed income effects by a slim margin, so that higher wages attracted a somewhat greater supply of labor. Specifically, a 10 percent rise in wages increased the hours of work of the Princeton student body by about 3 percent.

AN APPLICATION: THE LABOR SUPPLY PUZZLE

Income and substitution effects play an even more important role in explaining the striking historical trends in labor supply. Throughout the first three-quarters of this century, real wages rose, as Figure 16-1 clearly shows. Yet labor asked for and received *reductions* in the length of the workday and workweek. At the beginning of the century, the standard workweek was 50 to 60 hours (with virtually no vacations). Since then labor hours have generally declined to an average workweek of about 35 hours.

But in the two most recent decades, as real wages have fallen, there has been an accompanying increase in the number of family members who leave the home

[4] Mary P. Hurley, "An Investigation of Employment among Princeton Undergraduates during the Academic Year," senior thesis submitted to the Department of Economics, May 1975.

each day to earn wages: In 1993 nearly 65 percent of American married-couple families had two or more earners, compared with only 40 percent in 1970.[5] (This is called a rise in *labor-force participation*.) And, in the last few years, there has been a rise in overtime work, that is, workers laboring more than the standard number of hours in their firms. Thus, reduced real wages appear to have induced people to increase the quantity of labor they supply.

Where has the common-sense view of the matter gone wrong? Why, as hourly wages rose for 75 years, did workers not sell more of the hours they had available instead of pressing for a shorter and shorter workweek? And why, in recent years, have they sold more of their labor time as real wage rates stopped rising?

To answer the question, recall that any wage increase sets in motion *both* a substitution effect *and* an income effect. If only the substitution effect operated, then rising wages would indeed cause people to work longer hours because the high price of leisure makes leisure less attractive. But this reasoning leaves out the income effect.

Rising wages enable the worker to provide for the family with fewer hours of work. So the worker can afford to purchase more leisure without a cut in living standards. Thus, the income effect of increasing wages induces workers to work fewer hours. Similarly, falling wages reduce the worker's income. To preserve the family's living standard, the worker must seek additional hours of work; and the worker's spouse may have to leave their children in day care and take a job.

Thus, it is the strong income effect of rising wages that may account for the fact that labor supply has responded in the "wrong" direction, with workers working ever-shorter hours as real wages rose and longer hours as wages fell.

THE DEMAND FOR LABOR AND THE DETERMINATION OF WAGES

There is not much we can say about the demand for labor that was not already said about the demand for inputs in general in earlier chapters. Workers are hired (primarily) by profit-maximizing firms, which hire the quantity of an input at which the input's price (the market wage) is equal to its marginal revenue product.[6] If the marginal revenue product is above the input's price, the firm can increase its profit by acquiring more of the input—either to produce more output or to substitute for some other input. And the reverse is true when the marginal revenue product of the input is less than its price. Thus, the derived demand and, consequently, the demand curve for labor are determined by labor's marginal revenue product. Such a demand curve is shown as curve *DD* in Figure 16-3. The figure also includes a supply curve, labeled *SS*, much like the one depicted in Figure 16-2.

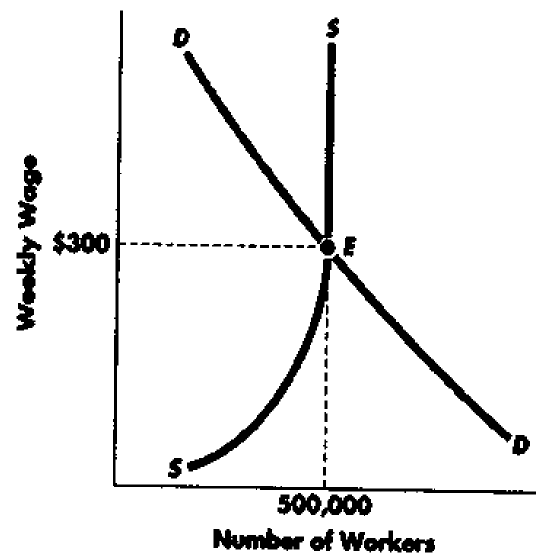
If there are no interferences with the operation of a free market in labor (such as minimum wages or unions—which we will consider later), equilibrium will be at point *E*, where the supply and demand curves intersect. In this example, 500,000 workers will be employed at a wage of $300 per week, so that the total income of the workers will be $300 × 0.5 million = $150 million.

---

[5]Bureau of the Census, *Statistical Abstract of the United States 1995*, 115th ed. (Washington, D.C.: U.S. Government Printing Office, 1995); and Bureau of the Census, *Historical Statistics of the United States, Colonial Times to 1970* (Washington, D.C.: U.S. Government Printing Office, 1975).

[6]To review, see Chapter 7, pages 152–156.

FIGURE 16-3 EQUILIBRIUM IN A COMPETITIVE LABOR MARKET



In a competitive labor market, equilibrium will be established at the wage that equates the quantity supplied with the quantity demanded. In this example, equilibrium is at point *E*, where demand curve *DD* crosses supply curve *SS*. The equilibrium wage is $300 per week and equilibrium employment is 500,000 workers.

WHY WAGES DIFFER

In reality, there is not one labor market but many—each with its own supply and demand curves and its own equilibrium wage. We all know that certain groups in our society (the young, the disadvantaged, the uneducated) earn relatively low wages and that some of our most severe social ills (poverty, crime, drug addiction) are related to this fact. But why are some wages so low while others are so high?

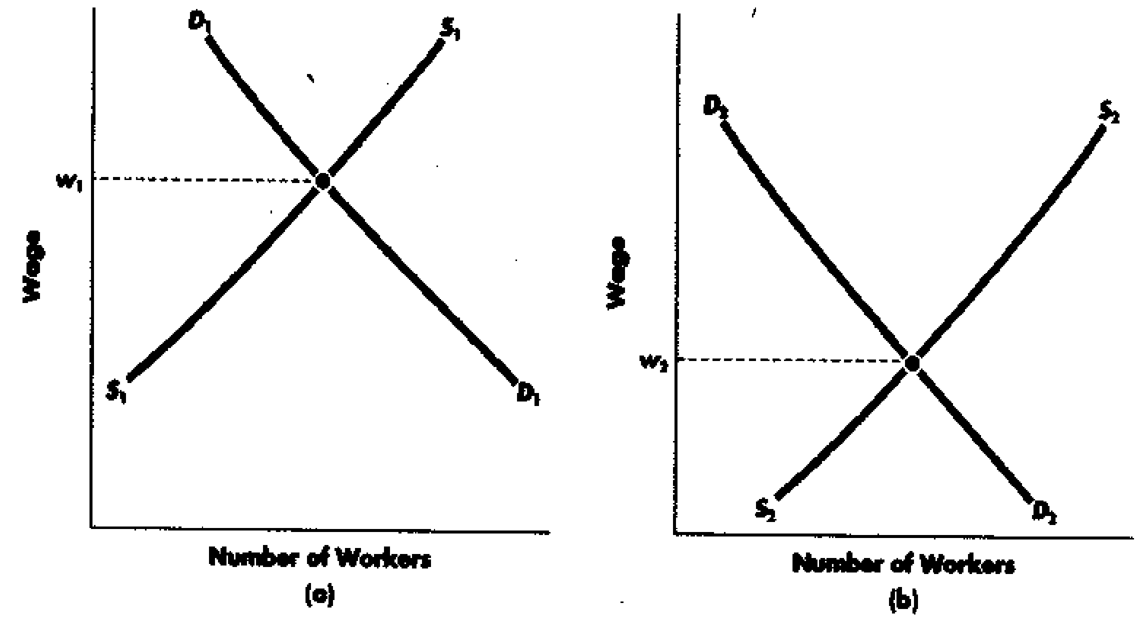
Supply-demand analysis at once tells us everything and nothing about this question. It implies that wages are relatively high in markets where demand is great and supply is small [see Figure 16-4(a)], while wages are comparatively low in markets where demand is weak and supply is high [see Figure 16-4(b)]. This is hardly startling news. To make the analysis useful, we must breathe some life into the supply and demand curves.

We start with demand. Why is the demand for labor greater in some markets than in others? Since the marginal revenue product of workers depends on their *marginal physical product* (MPP), variables that influence MPP will influence wages. A worker's marginal physical product depends, of course, on his own *abilities* and *degree of effort* on the job. But sometimes these characteristics are less important than the *other factors of production* that he has to work with. Workers in American industry are more productive than those in many other countries at least partly because they have generous supplies of machinery, natural resources, and technical know-how with which to work. As a consequence, they earn high wages. In other words, the marginal product of labor is raised by an abundance of efficient machinery and other inputs that increase the worker's effectiveness.

The marginal product of workers can also be increased by education, training, and experience (often referred to as *on-the-job training*). We will go into greater detail later about the role of education in wage determination.

Turning next to differences in the supply of labor to different areas, industries, or occupations, it is clear that the *size of the available working population* relative

FIGURE 16-4 WAGE DIFFERENTIALS



Panel (a): The market depicted here has a high equilibrium wage (w_1), because demand is high relative to supply. This can occur if qualified workers are scarce, or if productivity on the job is high or if the demand for the product is great. Panel (b): By contrast, the equilibrium wage is low here (w_2), where supply is high relative to demand. This can result from an abundant supply of qualified workers, low productivity, or weak demand for the product.

to the magnitude of industrial activity in a given area is of major importance. This helps explain why wages rose so high in sparsely populated Alaska when the Alaskan oil pipeline created many new jobs, and why wages have been and remain so low in Appalachia, where industry is dormant.

Second, it is clear that the *nonmonetary attractiveness* of any job will also influence the supply of workers to it. (The monetary attractiveness is the wage itself, which governs movements *along* the supply curve.) Jobs that people find pleasant and satisfying—such as teaching in suburban schools—will attract a large supply of labor and will consequently pay a low wage. In contrast, a premium will have to be paid to attract workers to jobs that are onerous, disagreeable, or dangerous—such as washing the windows of skyscrapers.

Finally, the amount of ability and training needed to enter a particular job or profession is relevant to its supply of labor. Brain surgeons and professional ice skaters earn generous incomes because there are few people as highly skilled as they, and because it is time consuming and expensive to acquire these skills even for those who have the ability.

ABILITY AND EARNINGS: THE RENT COMPONENT OF WAGES

In considering the effects of ability on earnings, it is useful to distinguish between skills that can be duplicated easily and skills that cannot. If Jill Jones has an ability that Sandra Smith cannot acquire, even if she undergoes extensive training, then the wages that Jones earns will contain an element of *economic rent,*

just as we saw in Chapter 15 to be true in the case of basketball star Shaquille O'Neal.[7]

Indeed, the salaries of professional athletes provide particularly clear examples of how economic rents can lead to huge wage differentials. Virtually anyone with moderate athletic ability can be taught to jump and shoot a basketball. But in most cases, no amount of training will teach the player to play basketball like Shaq. His high salary is a reward for his unique ability.

But many of the abilities that the market rewards generously—such as those of doctors and lawyers—clearly can be duplicated. Here the theory of rent does not apply, and we need a different explanation of the high wages that these skilled professionals earn. Once again, however, part of our analysis from Chapter 15 finds an immediate application because the acquisition of skills, through formal education and other forms of training, has much in common with business investment decisions. Why? Because the decision to gain more education in the hope of increasing future earnings involves a sacrifice of *current* income for the sake of *future* gain—precisely the hallmark of an investment decision.

INVESTMENT IN HUMAN CAPITAL

The idea that education is an investment is likely to be familiar even to students who have never thought explicitly about it. You made a conscious decision to go to college rather than to enter the labor market, and you are probably acutely aware that this decision is now costing you money—lots of money. Your tuition payments may be only a minor part of the total cost of going to college. Think of a high school friend who chose not to go to college and is now working. The salary that he or she is earning could, perhaps, have been yours. You are deliberately giving up this possible income in order to acquire more education.

In this sense, your education is an *investment* in yourself—a *human investment*. Like a firm that devotes some of its money to building a plant that will yield profits at some future date, you are investing in your own future, hoping that your college education will help you earn more than your high school-educated friend or enable you to find a more pleasant or prestigious job when you graduate. Economists call activities such as going to college *investments in human capital* because such activities give the human being many of the attributes of a capital investment.

Doctors and lawyers earn such high salaries partly because of their many years of training. That is, part of their wages can be construed as a *return on their (educational) investments*, rather than as economic rent. Unlike the case of Shaquille O'Neal, there are a number of people who conceivably *could* become surgeons if they found the job sufficiently attractive to endure the long years of training that are required. Few, however, are willing to make such large investments of their own time, money, and energy. Consequently, the few who do become surgeons earn very generous incomes.

Economists have devoted quite a bit of attention to the acquisition of skills through human investment. There is an entire branch of economic theory—called *human capital theory*—that analyzes an individual's decisions about education and training in exactly the same way as we analyzed a firm's decision to buy a machine or build a factory in the previous chapter. Though educational

[7]See the previous chapter, pages 365–368.

decisions can be influenced by love of learning, desire for prestige, and a variety of other factors, human capital theorists find it useful to analyze a schooling decision as if it were made purely as a business plan. The optimal length of education, from this point of view, is to stay in school until the marginal revenue (in the form of increased future income) of an additional year of schooling is exactly equal to the marginal cost.

One implication of human capital theory is that college graduates should earn substantially more than high school graduates to compensate them for their extra investments in schooling. Do they? Will your college investment pay off? Many generations of college students have supposed that it would, and recent data strongly confirm they were right. Indeed, the gap between the wages of workers with a college degree and those with a high school education has been widening: In 1973 the ratio of real (inflation-adjusted) average hourly wages of college grads to those of high school grads was about 1.45 (or about $17.50 per hour versus $12.00 per hour, measured in 1994 dollars); by 1994 that ratio had increased to 1.65 (or about $16.50 per hour versus $10.00 per hour). In other words, in 1973 college graduates earned 45 percent more than high school graduates; by 1994, they earned 65 percent more.[8]

The large income differentials earned by college graduates provide an excellent "return" on the tuition payments and sacrificed earnings that they "invested" while in school.

Human capital theory emphasizes that jobs that require more education *must* pay higher wages if they are to attract enough workers, because people insist on a financial return on their human investments. But the theory does not address the other side of the question: What is it about more educated people that makes firms willing to pay them higher wages? Put differently, the theory explains why the supply of educated people is limited but does not explain why the *demand* is substantial even at high wages.

Most human capital theorists complete their analyses by assuming that students in high schools and colleges acquire particular skills that are productive in the marketplace, thereby raising the marginal revenue products of those workers. In this view, educational institutions are factories that take less productive workers as their raw materials, apply doses of training, and produce more productive workers as outputs. This view of what happens in schools makes educators happy and accords well with common sense. However, a number of social scientists doubt that this is quite how schooling raises earning power.

EDUCATION AND EARNINGS: DISSENTING VIEWS

Just why is it that jobs with stiffer educational requirements typically pay higher wages? The common-sense view that educating people makes them more productive is not universally accepted.

EDUCATION AS A SORTING MECHANISM

One alternative view denies that the educational process teaches students anything directly relevant to their subsequent performance on jobs. In this view,

[8]"Winners Take All: Improved Competitiveness Exacts Social Price," *The Washington Post*, November 12, 1995, p. A14, which cites Bureau of the Census, *Current Population Survey*.

people differ in ability when they enter the school system and differ in more or less the same way when they leave. What the educational system does, according to this theory, is to *sort* individuals by ability. Skills like intelligence and self-discipline that lead to success in schools, it is argued, are closely related to the skills that lead to success in jobs. As a result, the abler individuals stay in school longer and perform better. Prospective employers know this, and consequently seek to hire those whom the school system has suggested will be the most productive workers.

THE RADICAL VIEW OF EDUCATION

Many radical economists question whether the educational system really sorts people according to ability. The rich, they note, are better situated to buy the best education and to keep their children in school regardless of ability. Thus, education may be one of the instruments by which a more privileged family passes its economic position on to its heirs while making it appear that there is a legitimate reason for firms to give them higher earnings. As radicals see it, education sorts people according to their social class, not according to their ability.

Radicals also hold a different idea about what happens inside schools to make workers more "productive." In this view, instead of serving primarily as instruments for the acquisition of knowledge and improved ability to think, what schools do primarily is teach people discipline—how to show up five days a week at 9 a.m., how to speak in turn and respectfully, and so on. These characteristics, radicals claim, are what business firms prefer and what causes them to seek more educated workers. They also suggest that the schools teach docility and acceptance of the capitalist status quo, and that this, too, makes schooling attractive to business.

THE DUAL LABOR MARKET THEORY

A third view of the linkages among education, ability, and earnings is part of a much broader theory of how the labor market operates—the theory of *dual labor markets*. Proponents of this theory suggest that there are two very different types of labor markets with relatively little mobility between them.

The "primary labor market" is where most of the economy's "good jobs" are—jobs like computer programming, business management, and skilled crafts that are interesting and offer considerable possibilities for career advancement. The educational system helps decide which individuals get assigned to the primary labor market and, for those who make it, greater educational achievement does indeed offer financial rewards.

The privileged workers who wind up in the primary labor market are offered opportunities for additional training on the job; they augment their skills by experience and by learning from their fellow workers; and they progress in successive steps to more responsible, better-paying positions. Where jobs in the primary labor market are concerned, dual labor market theorists agree with human capital theorists that education really is productive. But they agree with the radicals that admission to the primary labor market depends in part on social position, and that firms probably care more about steady work habits and punctuality than about reading, writing, and arithmetic.

Everything is quite different in the "secondary labor market"—where we find all the "bad jobs." Jobs like domestic and fast-food service, which are often the only ones inner-city residents can find, offer low pay, few fringe benefits, and virtually no training to improve the workers' skills. They are dead-end jobs with

little or no hope for promotion or advancement. As a result, lateness, absenteeism, and pilferage are expected as a matter of course, so that workers in the secondary labor market tend to develop the bad work habits that confirm the prejudices of those who assigned them to inferior jobs in the first place.

In the secondary labor market, increased education leads neither to higher wages nor to increased protection from unemployment—benefits generally offered elsewhere in the labor market. For this reason, workers in the secondary market have little incentive to invest in education.

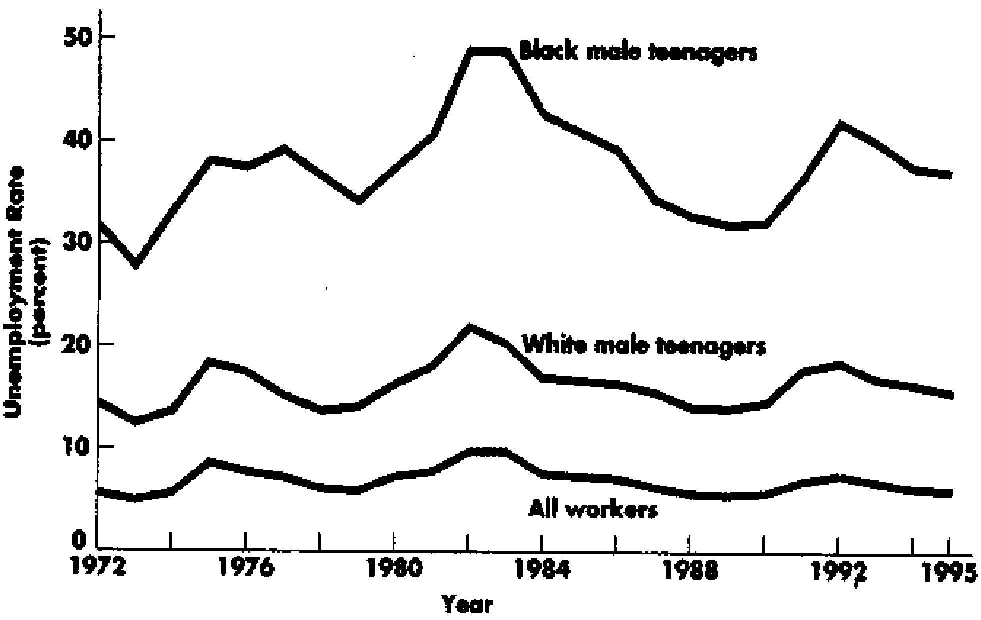
In sum, we have a well-established fact—that people with more education generally earn higher wages—but very little agreement on what accounts for this fact. Probably, there is some truth to all of the proposed explanations.

THE EFFECTS OF MINIMUM WAGE LEGISLATION

As we have observed, the "labor market" is really composed of many submarkets for labor of different types, each with its own supply and demand curves. One particular labor market always seems to have higher unemployment than the labor force as a whole: the job market for teenagers.

Figure 16-5 shows the record. It indicates that whenever the unemployment rate for all workers goes up or down, the teenage unemployment rate almost always moves in the same direction, but more dramatically. Thus, when things are generally bad, things are much, much worse for teenage workers, and

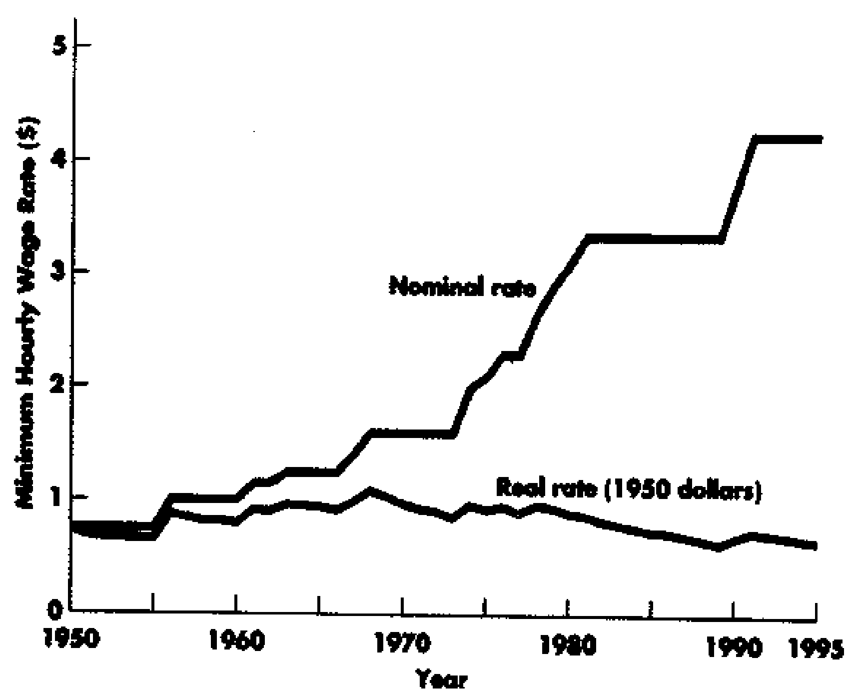
FIGURE 16-5 THE TEENAGE UNEMPLOYMENT PROBLEM



Teenage unemployment rates have consistently been much higher than the overall unemployment rate, and black teenagers have fared worse than white teenagers. For the most part the three employment rates have moved up and down together, as this chart shows. Note: A teenager here is defined as a person aged 16 to 19 years.

SOURCE: *Economic Report of the President* (Washington, D.C.: U.S. Government Printing Office, various years).

FIGURE 16-6 THE MINIMUM WAGE, 1950–1995



After 45 years of sporadic increases in the federal minimum hourly wage rate, the real (inflation-adjusted) rate has remained just about the same.

Note: Nominal rates deflated by CPI (1950 = 100).

SOURCE: Bureau of the Census, *Statistical Abstract of the United States 1995*, 115th ed. (Washington, D.C.: U.S. Government Printing Office, 1995).

especially for black teenage workers. Despite social and legislative pressures against race discrimination, efforts to improve the quality of education available to children in the inner cities, and many related programs, there has been no relative improvement in black teenage unemployment in recent years.

One reason is that teenagers generally have not completed their educations and have little job experience, hence their marginal revenue products tend to be relatively low. Until recently, many economists argued that this fact, together with minimum wage laws that prevent teenagers from accepting wages commensurate with their low marginal revenue products, is the main cause of high teenage unemployment. The reasoning is that legally imposed high wages make it too expensive to hire teenagers. Recent studies of the data suggest, however, that a rise in minimum wage produces little if any cut in demand for teen labor.

We should also note that inflation has eaten into the real value of the minimum wage over the years. Figure 16-6 shows clearly that after 45 years of sporadic increases in the nominal minimum wage rate, the real (inflation-adjusted) rate has remained virtually unchanged.

UNIONS AND COLLECTIVE BARGAINING

Our analysis of competitive labor markets has ignored one rather distinctive feature of the markets for labor: The supply of labor is not at all competitive in many labor markets; instead it is controlled by a labor monopoly, a union.

While significant, unions in America are not nearly so important as is popularly supposed. For example, most people who are unfamiliar with the data are astonished to learn that only about 15 percent of American workers belong to unions. This percentage is much higher than it was before the New Deal, when

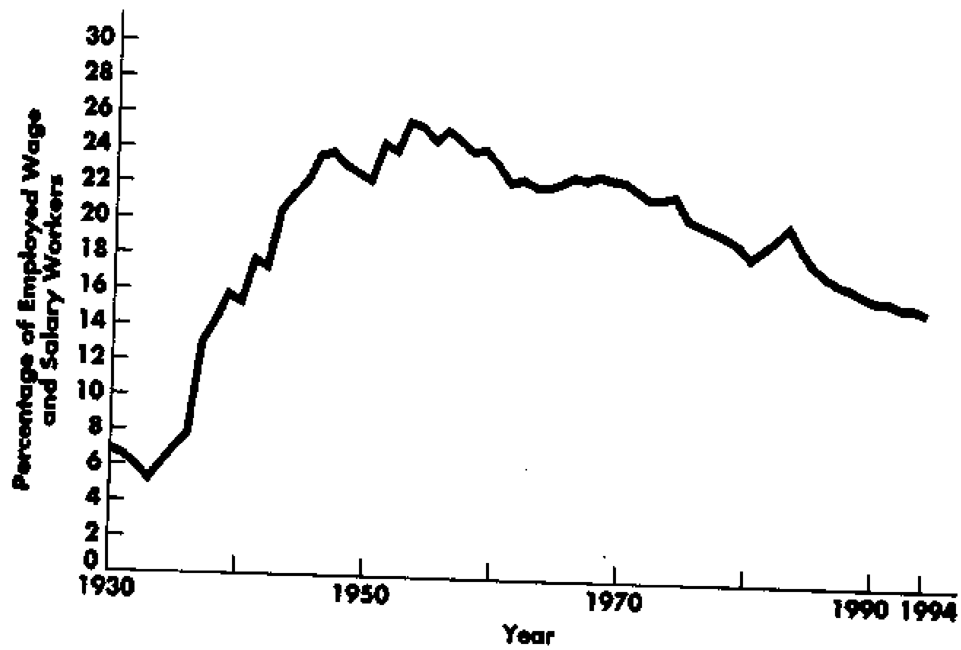
unions were quite unimportant in this country, but lower than it was in the heyday of unionism in the mid-1950s, when the figure was about 25 percent (see Figure 16-7). Since then, the unionization rate has fallen with few interruptions. This trend appeared to be reversing in 1993–1994 when union membership increased by 3 percent, ending a 14-year decline. But since the number of wage and salary workers grew at an even faster rate during the same period, the overall share of unionized jobs continued to fall.

Why has the extent of unionization in America been declining? One reason is the shift of the U.S. labor force (like that of every other industrial country) into service industries and out of manufacturing, where unions traditionally had their base. Deregulation forced airlines and trucking companies to compete more intensely, and it may also have influenced other firms to hire less expensive nonunion labor.

In addition, there seems to have been a shift in the preferences of American workers away from unions. The increasing share of women in the labor force may have contributed to this trend, since women have traditionally been less prone than men to join unions. But there is evidence that women's attitudes toward unions may be changing. Since 1983, the percentage of women in the unionized work force has risen from 33 percent to 39 percent.

Finally, American unions have been under increasing pressure in the 1990s owing to stronger competition both at home and from abroad. In response, firm after firm has closed plants and eliminated jobs. The downsizing trend has made it even more difficult for unions to win concessions that improve the economic

FIGURE 16-7 UNIONIZATION IN THE UNITED STATES, 1930–1994



In 1930, unions had enrolled just under 7 percent of the U.S. labor force, and by 1933 this figure had slipped to barely above 5 percent. Unionization took off with FDR's New Deal, reaching almost 16 percent of the labor force by 1939. It then drifted irregularly upward to a peak of about 25 to 26 percent of all workers in the mid-1950s, from which it has since fallen more or less steadily back to 15.5 percent in 1994.

SOURCE: Bureau of Labor Statistics, *Employment and Earnings* (Washington, D.C.: U.S. Government Printing Office, January issues, various years).

positions of their members. And that, in turn, has reduced the attractiveness of union membership.

Unionization is also much less prevalent in America than it is in most other industrialized countries. For example, 32 percent of German workers and 83 percent of Swedish workers belong to unions.[9] The differences are quite striking and doubtless have something to do with our tradition of "rugged individualism."

The main sector of the U.S. economy in which the unions are still fairly healthy is government employment. City, state, and federal employees are relatively unionized (about 39 percent of public sector workers are union members), perhaps because their job opportunities are controlled by political forces, and politicians find it far harder to resist the unions than private firms do.[10] However, here too, the unions are under siege as more and more government services are contracted out to private businesses. For example, the state of New Jersey used the threat of turning highway toll collection over to private enterprise as an effective bargaining chip in 1995 negotiations with the toll-collectors' union.

THE DEVELOPMENT OF UNIONISM IN AMERICA

Serious unionism in America is not much more than a century old. Large-scale unions in this country began with the Knights of Labor—a politically oriented organization quite different from the unions of today. Toward the end of the 19th century its membership approached 750,000; but it obtained neither higher wages nor improved working conditions, and so it declined rapidly.

The current union structure began to take shape in 1881 with the founding of the American Federation of Labor (AFL) by Samuel Gompers. Working conditions then were incredibly bad by today's standards, and unscrupulous practices by many employers fostered the growth of unions (see the box on page 391).

Gompers strongly believed that unions should be nonpolitical organizations seeking more pay, better working conditions, longer vacations, and so on. He also believed that unions should be organized along craft lines—carpenters in one union, plumbers in another—rather than trying to include all types of workers within a given industry.

The AFL grew steadily from about 1900 until the 1920s, went into decline during the Roaring Twenties, but then grew rapidly thanks to the favorable legislation of the Roosevelt administration in the 1930s. The Norris–La Guardia Act of 1932 sharply limited the power of the federal courts to interfere in labor disputes. In 1938, the Fair Labor Standards Act abolished child labor, imposed a minimum wage on most activities in interstate commerce, and required extra pay for overtime work.

Even more important, the National Labor Relations Act (Wagner Act) in 1935 guaranteed workers the right to form unions and to choose unions to represent them in collective bargaining. It also set up the National Labor Relations Board (NLRB) to protect labor from "unfair labor practices" by employers. Today the NLRB oversees elections in firms to determine which union will represent the workers. It can also force employers to take back workers who it considers to have been fired unjustly.

---

[9]Organization for Economic Cooperation and Development, *Employment Outlook, 1994* (Paris: OECD, July 1994), p. 173.

[10]Bureau of the Census, *Statistical Abstract of the United States 1995*, 115th ed. (Washington, D.C.: U.S. Government Printing Office, 1995).

THE WAY IT WAS

The calamitous Triangle Shirtwaist Factory fire of 1911, in which 146 women and girls lost their lives, was a landmark in American labor history. It galvanized public opinion behind the movement to improve conditions, hours, and wages in the sweatshops. Pauline Newman went to work in the factory, located in New York City's Lower East Side, at the age of eight. Many of her friends lost their lives in the fire. She went on to become an organizer and executive of the newly formed International Ladies Garment Workers' Union. In her words:

> We started work at seven-thirty in the morning, and during the busy season we worked until nine in the evening. They didn't pay you any overtime and they didn't give you anything for supper money. . . .
>
> The employers didn't recognize anyone working for them as a human being. You were not allowed to sing. . . . We weren't allowed to talk to each other. . . . If you went to the toilet and you were there longer than the floor lady thought you should be, you would be laid off for half a day and sent home. And, of course, that meant no pay. You were not allowed to have your lunch on the fire escape in the summertime. The door was locked to keep us in. That's why so many people were trapped when the fire broke out. . . .
>
> You were expected to work every day if they needed you and the pay was the same whether you worked extra or not.
>
> Conditions were dreadful in those days. We didn't have anything. . . . There was no welfare, no pension, no unemployment insurance. There was nothing. . . . There was so much feeling against unions then. The judges, when one of our girls came before him, said to her: "You're not striking against your employer, you know, young lady. You're striking against God," and sentenced her to two weeks.
>
> I wasn't at the Triangle Shirtwaist Factory when the fire broke out, but a lot of my friends were. . . . The thing that bothered me was the employers got a lawyer. How anyone could have *defended* them!—because I'm quite sure that the fire was planned for insurance purposes. And no one is going to convince me otherwise. And when they testified that the door to the fire escape was open, it was a lie! It was never open. Locked all the time. One hundred and forty-six people sacrificed, and the judge fined Blank and Harris seventy-five dollars!

SEQUEL: THE PROBLEM CONTINUES

> Nakhon Pathom, Thailand—One moment they were mothers and fathers of Thailand, sewing and stuffing Playskool "Water Pets" for daughters and sons of the United States.
>
> A moment later they were trapped in a fire fueled by the fabric of toys meant to comfort and amuse.
>
> They listened for a fire alarm that never rang. They searched for sprinklers and fire hoses that had never been installed. They ran for doors that were locked to prevent theft.
>
> They knelt in prayer and waited in pain. Then they died.
>
> The fire . . . at the Kader Industrial Co. toy factory here [in 1993] killed at least 188 workers, almost all young women. . . .
>
> The fire was reminiscent of the infamous 1911 blaze that claimed 146 lives at the Triangle Shirtwaist Co. in New York.

SOURCES: Excerpted from Joan Morrison and Charlotte Fox Zabusky, *American Mosaic: The Immigrant Experience in the Words of Those Who Lived It*. 1980. Reprinted by permission of the publisher, E. P. Dutton, Inc.; and Mitchell Zuckoff, "Trapped by Poverty, Killed by Neglect," *The Boston Globe*, July 10, 1994, p. 19.



The Triangle Factory, now part of New York University.

An **INDUSTRIAL UNION** represents all types of workers in a single industry, such as auto manufacturing or coal mining.

A **CRAFT UNION** represents a particular type of skilled worker, such as plumbers or electricians, regardless of what industry they work in.

By no coincidence, in the year of the Wagner Act, John L. Lewis founded the Congress of Industrial Organizations (CIO), a federation of many **industrial unions** that at first rivaled the AFL for leadership of the U.S. labor movement. Those who advocated industrial unions believed that many specialized **craft unions** (which often quarreled among themselves) were not likely to be very powerful in their dealings with large employers. Despite their differences, the AFL, with its craft unions, and the CIO, with its industrial unions, eventually merged in 1955.

The favorable public attitude toward unions soured somewhat after World War II, perhaps because of the rash of strikes that took place in 1946. One result of these strikes was the Taft-Hartley Act of 1947, which specified and outlawed certain "unfair labor practices" by unions and which sought to shift some of

A **CLOSED SHOP** is an arrangement that permits only union members to be hired.

A **UNION SHOP** is an arrangement under which nonunion workers may be hired, but then must join the union within a specified period of time.

labor's power back to management. Specifically, the act severely limited **closed shops** that hire only union members; permitted state governments to ban the **union shop,** an arrangement that *requires* employees to join the union; and provided for court injunctions to delay strikes that threaten the national interest for an 80-day "cooling-off" period.

Today, the character of American unionism is still somewhat unsettled. Unions are struggling very hard to make inroads into labor markets that by tradition have not been unionized—such as the agricultural and white-collar office markets. They have achieved notable successes in organizing teachers, government employees, and many other types of white-collar workers. Between 1984 and 1994, the number of management, executive, and professional workers belonging to unions increased by 1.6 million.

U.S. labor unions are very different from those in Europe and Japan. For one thing, U.S. unions are strongly committed to capitalism and have rarely espoused socialism—unlike their western European counterparts. American unions also often see themselves as adversaries of management, unlike Japanese unions. This country's century-long tradition of hostile labor-management relations has impeded current attempts to emulate the Japanese model of labor-management cooperation. However, several U.S. plants that are owned and run by Japanese companies reportedly have achieved an unprecedented degree of trust and support between the labor force and management. There are other cases though in which such firms have had extremely poor labor relations. In general, U.S. labor still seems to feel that American employers are all too likely to adopt unfair practices unless they are restrained by powerful unions.

UNIONS AS LABOR MONOPOLIES

Unions require that we alter our economic analysis of the labor market in much the same way that monopolies required us to alter our analysis of the goods market (see Chapter 11). Recall that a monopoly seller of goods selects the point on its demand curve that maximizes its profits. Much the same idea applies to a union, which is, after all, a monopoly seller of labor. It too faces a demand curve—derived this time from the marginal revenue product schedules of firms—and can choose the point on that curve that suits it best.

The problem for the economist trying to analyze union behavior—and perhaps also for the union leader trying to select a course of action—is how to decide which point on the demand curve is "best." There is no obvious single goal analogous to profit maximization that clearly delineates what a union should do. Instead there are a number of *alternative* goals that sound plausible.

ALTERNATIVE UNION GOALS

The union leadership may, for example, decide that the size of the union is pretty well fixed and try to force employers to pay the highest wage they will pay without firing any of the union members. But this is a high-risk strategy for a union. Firms forced to pay such high wages will be at a competitive disadvantage compared with firms that have nonunion labor, and they may even be forced to shut down. Alternatively, union leaders may assign priority to increasing the size of

their union. They might even try to make employment as large as possible by accepting a wage just above the competitive level. One way, but certainly not the only way, to strike a balance between the conflicting goals of maximizing wages and maximizing employment is to maximize the product of the two—which is the total earnings of all workers taken together.

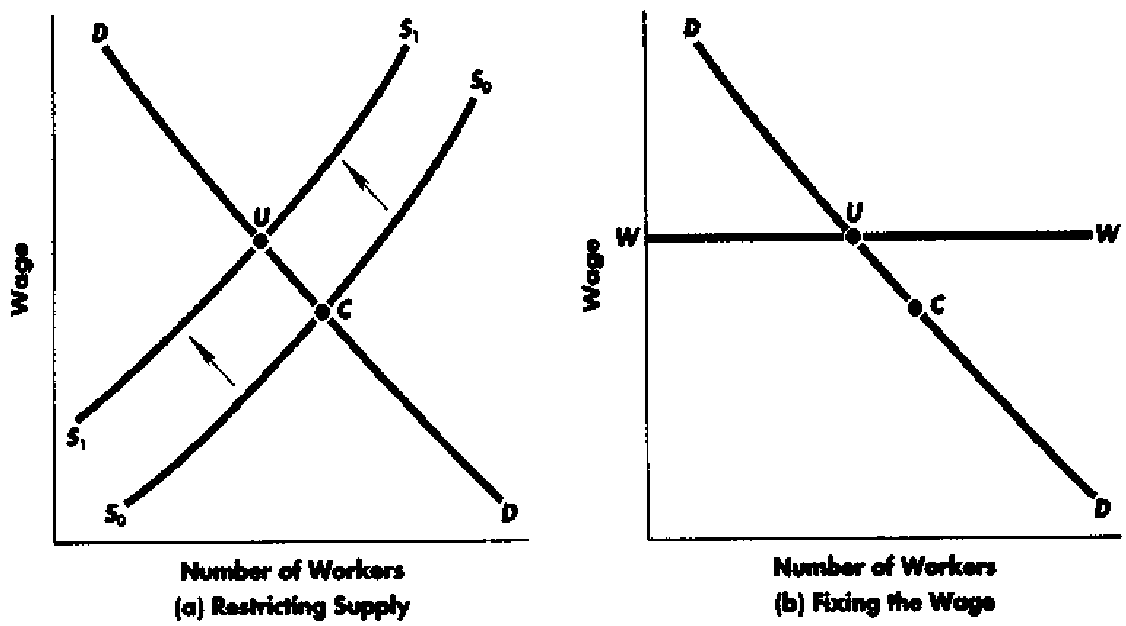
The basic conclusion of these alternative possible goals for unions is this: Even if unions, as monopoly sellers of labor, have the power to push wages above the competitive level, such increases in wages normally can be achieved only by reducing the number of jobs, because the demand curve for labor is downward sloping. Just as monopolists must limit their outputs to push up their prices, so must unions restrict employment to push up wages.

ALTERNATIVE UNION STRATEGIES

How would a union that has decided to push wages above the competitive level accomplish this task? Figure 16-8 illustrates two principal strategies. There we let point *U* on demand curve *DD* be the union's choice, and point *C* the competitive equilibrium.

In Figure 16-8(a), we suppose that the union pursues its goal by *restricting supply*. By keeping out some workers who would like to enter the industry or occupation, it shifts the supply curve of labor inward from S_0S_0 to S_1S_1. Craft unions often accomplish this by requiring long apprenticeships. Such unions sometimes offer only a small number of new memberships each year, largely to replace members who have died or retired. Membership in such a union is extremely

FIGURE 16-8 TWO UNION STRATEGIES



The two graphs indicate two alternative ways for the union to move from point *C* to point *U*. In Panel (a), it keeps some workers out of the industry, thereby moving the supply curve to the left from S_0S_0 to S_1S_1. As a consequence, wages rise. In Panel (b), it fixes a high wage (*W*) and provides labor only at this wage. Therefore, firms reduce employment. The effects are the same under both strategies.

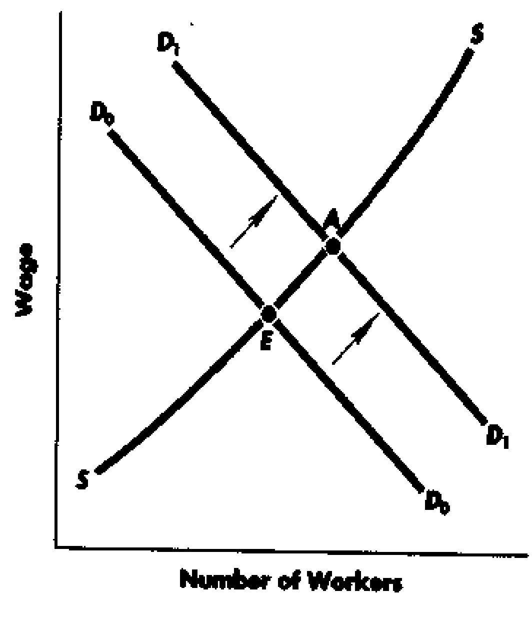
valuable and is sometimes offered primarily to children of current members. If all or almost all of those members are and have been white, that is also a roundabout but effective way to close the door to union membership and jobs to minority groups.

In Figure 16-8(b), instead of restricting supply, the union simply *sets a high wage rate,* *W* in the example. Then it is the employers who restrict entry into the job, because with wages so high they will not want to employ many workers. This second strategy is more typically employed by industrial unions like the United Automobile Workers or the United Mine Workers. As the figure makes clear, the two wage-raising strategies achieve the same result (point *U* in either case) by what turns out to be the same means. Wages are raised only by reducing employment in either case.

In some exceptional cases, however, a union may be able to achieve wage gains without sacrificing employment. To do this, the union must be able to exercise effective control over the demand curve for labor. Figure 16-9 illustrates such a possibility. Union actions push the demand curve outward from D_0D_0 to D_1D_1, simultaneously raising both wages and employment. Typically, this is difficult to do. One way to do it is by *featherbedding*—forcing management to employ more workers than they really need.[11] Quite the opposite technique is to institute a campaign to raise worker productivity, which some unions seem to have been able to do. Alternatively, the union can try to raise the demand for the company's

FIGURE 16-9 UNION CONTROL OVER THE DEMAND CURVE

This diagram indicates yet a third way in which unions may affect the labor market—a pleasant alternative for workers in that wages can be raised while adding to employment. Strong unions may succeed in raising the demand curve from D_0D_0 to D_1D_1 by featherbedding, raising worker productivity, or using their influence to increase demand for the product. Equilibrium then shifts from point *E* to point *A*.



[11]The best-known example of featherbedding involved the railroad unions, which for years forced management to keep "firemen" in the cabs of diesel engines, in which there were no burning fires. Similarly, the musicians' union in New York City forces Broadway producers who use certain theaters to employ a minimum number of musicians—whether or not they actually play music. Of course, it is not only unionized labor that has tried to create an artificial demand for its services. Lawyers, doctors, and business firms, among others, have sought ways to induce consumers to buy more of their products and services. EXERCISE: Can you think of ways in which they have done this?

product either by flexing its political muscle (for example, by obtaining legislation to reduce foreign competition) or by appealing to the public to buy union products.

HAVE UNIONS REALLY RAISED WAGES?

To what extent do union members actually earn higher wages than nonmembers? The consensus would probably surprise most people. Economists estimate that most union members' wages are about 15 percent above those of nonmembers who are otherwise identical (in skills, geographical locations, and so on). While certainly not negligible, and while there are indications that this number has been going upward slightly, this can hardly be considered a huge differential. Narrowing the gap even further is the fact that nonunion workers do not have to pay the dues that are required of union members.

This 15 percent differential does not mean, however, that unions have raised wages no more than 15 percent. Some observers believe that union activity has also raised wages of nonunion workers by forcing nonunion employers to compete harder for their workers. If so, the differential between union and nonunion workers will be less than the amount by which unions raised wages overall.

MONOPSONY AND BILATERAL MONOPOLY

Our analysis thus far oversimplifies matters in several important respects. For one thing, it envisions a market situation in which one powerful union is dealing with many powerless employers: We assume the labor market is monopolized on the selling side but competitive on the buying side. There are industries that more or less fit this model. The giant Teamsters' union negotiates with a trucking industry consisting of thousands of firms, most of them quite small and powerless. Similarly, most of the unions within the construction industry are much larger than the firms with which they bargain.

But there are many cases that simply do not fit the model. The "Big Three" automakers do not stand idly by while the UAW picks its favorite point on the demand curve for auto workers. Nor does the Steelworkers' union sit across the bargaining table from representatives of a perfectly competitive industry. In these and other industries, while the union certainly has a good deal of monopoly power over labor supply, the firms also have some **monopsony** power over labor demand. This means that the firms may deliberately reduce the quantity of labor they demand as a way to force down the equilibrium level of wages. We can calculate the profit-maximizing restriction in the quantity of labor the same way we determined a monopolist's profit-maximizing restriction of output in Chapter 11.

MONOPSONY refers to a market situation in which there is only one buyer.

Analysts find it difficult to predict the wage and employment decisions that will emerge when both the buying and selling side of a market are monopolized—a situation called **bilateral monopoly**. The difficulties here are similar to those we encountered in considering the behavior of oligopolistic industries in Chapter 12. Just as one oligopolist is acutely aware that rivals are likely to react to anything he does, a union dealing with a monopsony employer knows that any move it makes will elicit a countermove by the firm. And this knowledge makes the first decision that much more complicated. In practice, the outcome of bilateral monopoly will depend partly on economic logic, partly on the relative

BILATERAL MONOPOLY is a market situation in which there is both a monopoly on the selling side and a monopsony on the buying side.

power of the union and management, partly on the skill and preparation of the negotiators, and partly on luck.

Still, we can be a bit more concrete about the outcome of the wage determination process under bilateral monopoly. A monopsonist employer unrestrained by a union will use its market power to force wages down below the competitive level, just as a monopoly seller uses its market power to force prices higher. It accomplishes this by reducing its demand for labor below what would otherwise be the profit-maximizing amount, thereby cutting both wages and the number of workers employed.

However, a union may be in a position to prevent this from happening. It can deliberately set a floor on wages, pledging its members not to work at all at any wage level below this floor. So, a union may force the monopsony employer to pay higher wages and, simultaneously, to hire more workers than he otherwise would.

Even though it is hard to think of industries that are pure monopsonists in their dealings with labor, these conclusions are of some importance in reality. For the fact is that large, oligopolistic firms do often engage in one-on-one wage bargaining with the unions of their employees, and there is reason to believe that the resulting bargaining process closely resembles the workings of the bilateral monopoly model that we just described.

COLLECTIVE BARGAINING AND STRIKES

The process by which unions and management settle upon the terms of a labor contract is called *collective bargaining*. Unfortunately, there is nothing as straightforward as a supply-demand diagram to tell us what wage level will emerge from a collective bargaining session.

Furthermore, actual collective bargaining sessions range over many more issues than wages. For example, fringe benefits such as pensions, health and life insurance, overtime pay, seniority privileges, and work conditions are often crucial issues. Many labor contracts specify in great detail the rights of labor and management to set work conditions—and also provide elaborate procedures for resolving grievances and disputes. This list could go on and on. The final contract that emerges from collective bargaining may well run to many pages of fine print.

With the issues so varied and complex, and with the stakes so high, it is no wonder that both labor and management employ skilled professionals who specialize in preparing for and carrying out these negotiations. The bargaining in these sessions is often heated, with outcomes riding as much on personalities and the skills of the negotiators as on cool-headed logic and economic facts. Negotiations may last well into the night, with each side making threats and seeming to try to wear the other out. Unions, for their part, generally threaten strikes or work slow-downs. Firms counter that they would rather face a strike than give in, or they may even threaten to close the plant without a strike. (This is called a *lock-out*.)

MEDIATION AND ARBITRATION

Where the public interest is seriously affected, or when the union and firm reach an impasse, government agencies may well send in a *mediator*, whose job is to try to speed up the negotiation process, as occurred during the 1994–1995 base-

ball strike. (See the box below.) As an impartial observer, the mediator sits down with both sides separately to discuss their problems and then tries to persuade each to make concessions. At some stage, when an agreement looks possible, she may call them back together for another bargaining session in her presence. Mediators, however, have no power to force a settlement. Their success hinges on their ability to smooth ruffled feathers and to find common ground.

Sometimes, in cases where unions and firms simply cannot agree and where neither wants a strike, differences are finally settled by *arbitration*—the appointment of an impartial individual empowered to settle the issues that negotiation could not resolve. This happens often, for example, in wage negotiations in professional sports or for municipal jobs such as police and firefighters. Most recently, a federal arbitrator was called in to resolve the labor dispute between American Airlines and its flight attendants' union. In fact, in some vital sectors where a strike is too injurious to the public interest, the labor contract or the law

THE 1994-1995 BASEBALL STRIKE

The major league baseball strike—the longest in major sports history—lasted 234 days. Players gave up $230 million in salaries. Owners claimed to have lost better than $700 million, with some red ink still to be tabulated. The dispute left the national pastime in shambles, and in the end it remained unclear whether any of the really significant management/labor issues were resolved.

Deciding who was to blame for the strike was no easy task, and both sides seemed to have valid arguments. Owners wanted a salary cap, considering that the average major-leaguer earns $1.2 million a year—20 times more than just two decades ago.

Another longstanding sticking point between players and management was baseball's antitrust exemption. This 73-year-old perk precludes the possibility of players suing owners, even if they suspect collective bargaining is being conducted in a monopolistic manner. A decision by owners to try to impose a salary cap collusively—much like other major sports—added critical momentum to the strike threat.

President Clinton declared it was time for the impasse to end when the strike approached its 6th month, when the two sides were no longer talking at all. Clinton recommended that arbitration replace mediation and clearly signaled to owners that their coveted antitrust exemption might be placed in peril. But Congress, which has exhibited an historic aversion to interfering in private labor disputes, defeated a bill to force binding arbitration in February 1995.

About all that was proven is that strikes are, at least in the case of baseball, an exceptionally unentertaining way to squander around $1 billion.

SOURCES: Mark Maske, "Baseball Strike Ends, Season Opens April 26," *The Washington Post*, April 2, 1995; Sam Donnellon, "On Anniversary of Strike, Players, Management Seem No Closer to Agreement," *The Philadelphia Daily News*, August 11, 1995; "An Empty Field of Dreams?" *Time*, August 8, 1994, p. 65; "Owners, Players Prepare to Spoil Summer," *U.S. News & World Report*, June 27, 1994, p. 17; "An Unwhole New Ball Game," *Time*, April 17, 1995, p. 48; "The Ninth Strike," *The Economist*, April 8, 1995, p. 29; and "Clinton Tries to Win One for the Babe," *U.S. News & World Report*, February 6, 1995, p. 12.

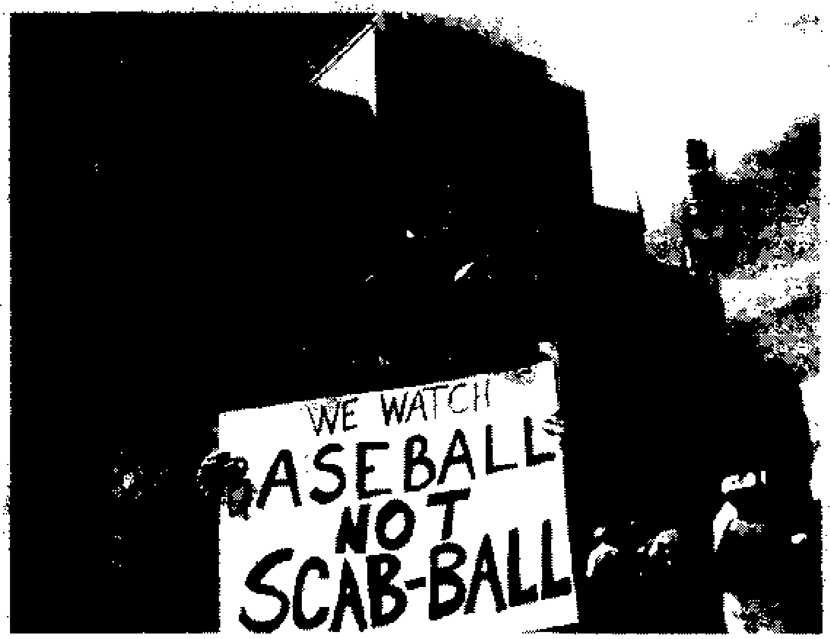
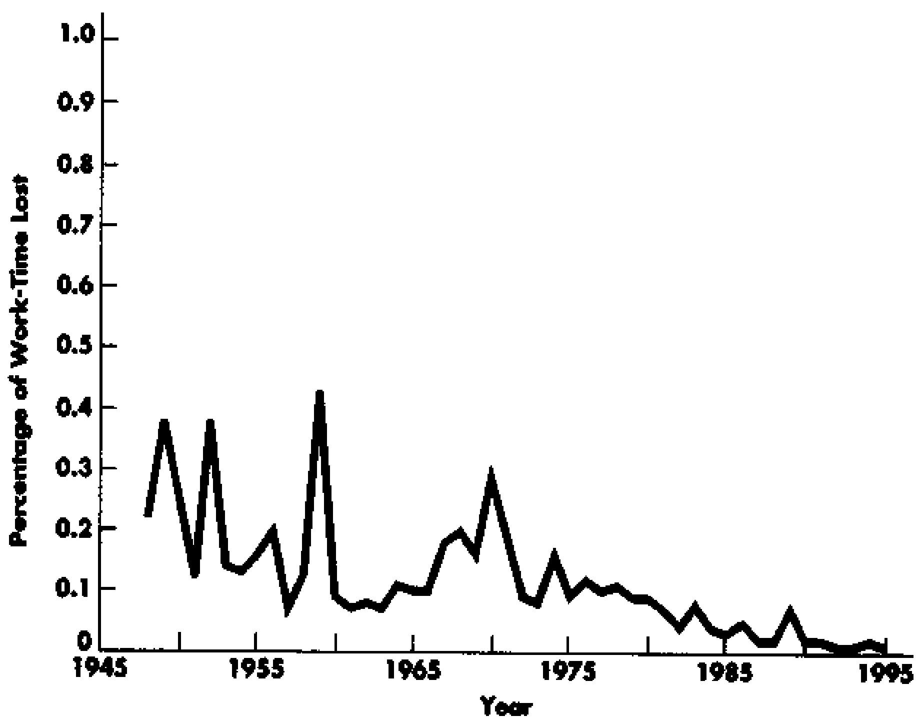


FIGURE 16-10 WORK-TIME LOST IN THE UNITED STATES BECAUSE OF STRIKES, 1948–1995



The fraction of total work-time lost to work stoppages varies greatly from year to year, but is never very large. It has been under one-tenth of 1 percent since the late 1970s. The worst year for strikes was 1946, and it is probably no coincidence that the Taft-Hartley law was enacted the following year.

SOURCE: Bureau of Labor Statistics, *Monthly Labor Review*, various issues.

may stipulate that there must be *compulsory arbitration* if the two parties cannot agree. However, both labor and management are normally reluctant to accept this procedure.

STRIKES

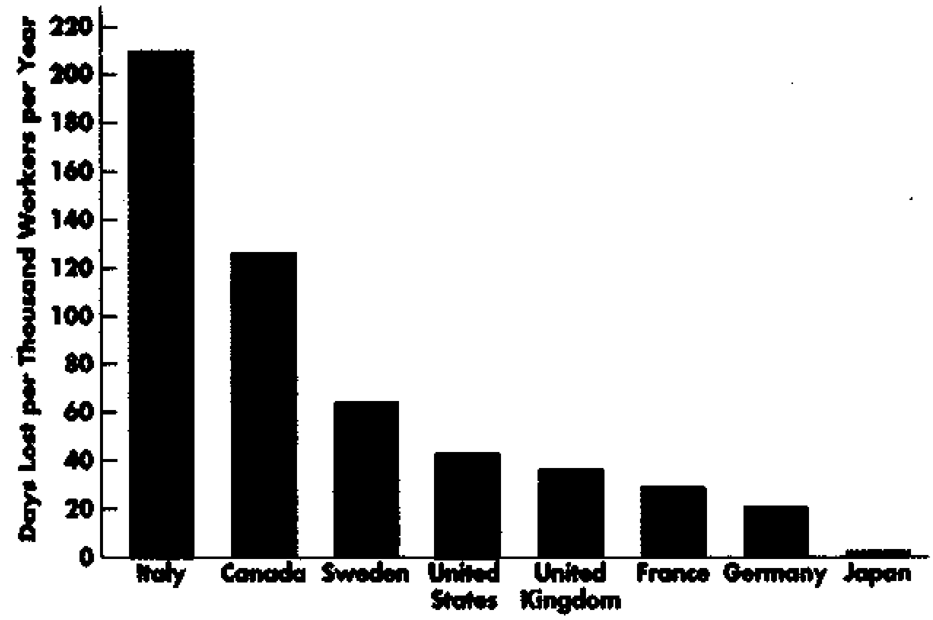
Most collective bargaining situations do not lead to strikes. But the right to strike, and to take a strike, remain fundamentally important for the bargaining process. Imagine, for example, a firm bargaining with a union that was prohibited from striking. It seems likely that the union's bargaining position would be quite weak. On the other hand, a firm that always capitulated rather than suffer a strike would be virtually at the mercy of the union. So strikes, or more precisely, the possibility of strikes, serve an important economic purpose.

Fortunately, however, the incidence of strikes is not nearly so common as many people believe. Figure 16-10 reports the percentage of work-time of labor lost as a result of strikes in the United States from 1948 to 1995. Despite the headline-grabbing nature of major national strikes, the total amount of work-time lost to strikes is truly trivial—far less, for example, than the time lost to coffee breaks! Compared with other nations, America suffers more from strikes than, say, Japan, but it has many fewer strikes than such countries as Italy and Canada (see Figure 16-11).

RECENT DEVELOPMENTS IN THE U.S. LABOR MARKET

The past decade has brought a number of noteworthy developments to the U.S. labor market. We have already discussed the decline in union membership and mentioned the increasing share of women in the labor force. The latter has given

FIGURE 16-11 THE INCIDENCE OF STRIKES IN EIGHT INDUSTRIAL COUNTRIES, 5-YEAR AVERAGES, 1990–1994



Although strikes are less common in the United States than they are in Italy or Canada, they are much more common here than in Japan. Note: Data are averages for the 5-year period 1990 to 1994, except for Germany and Sweden, for which the graph reports 4-year averages for 1990 to 1993.

SOURCE: Bureau of Labor Statistics, Office of Productivity and Technology, *Industrial Disputes, Workers Involved, and Worktime Lost, 15 Countries, 1955–1994,* unpublished data, August 1995.

rise to a number of associated issues such as the extent to which employers or the government should be expected to provide for child care. (For example, day-care centers are needed to make it possible for mothers to work, thereby avoiding damaging breaks in their careers.) It was also a driving force behind enactment of the Family and Medical Leave Act in 1993. The law enables workers of businesses with 50 or more employees to take up to 12 weeks of unpaid leave for such things as the birth of a child or the serious illness of a family member. Most employees are guaranteed their same positions upon returning to work.

IMMIGRANT WORKERS

The role of immigrants has also received a good deal of attention; indeed, it has become a major political issue of late. Foreign workers, many of whom immigrated illegally, do much of the relatively unskilled work in the United States. Crop picking in Florida and sewing in sweatshops in New York's Chinatown are two well-known examples; and many American enterprises fear what would happen if they were deprived of this labor force. But American workers are concerned about competition from this source. In a sign of the growing opposition to illegal immigration, California voters overwhelmingly approved Proposition 187 in 1994, a ballot initiative denying most public services to illegal aliens. The state debate spilled over into the national arena, prompting Congress to consider a wide range of immigration reform legislation, including a proposal for a national ID system aimed at preventing illegal immigrants from getting jobs in the United States.

GROWTH IN LONG-TERM UNEMPLOYMENT

Another very disturbing development in the labor market is the rising proportion of the unemployed who are out of work for protracted periods. This

phenomenon has plagued every leading industrial country, as Figure 16-12 illustrates. The graph shows that over the 18-year period from 1975 through 1993 the share of unemployed people in the United States who had been jobless 1 year or more grew 74 percent. But this increase was small compared to the corresponding increase in European countries, with long-term unemployment rising 772 percent in France and 687 percent in Sweden.

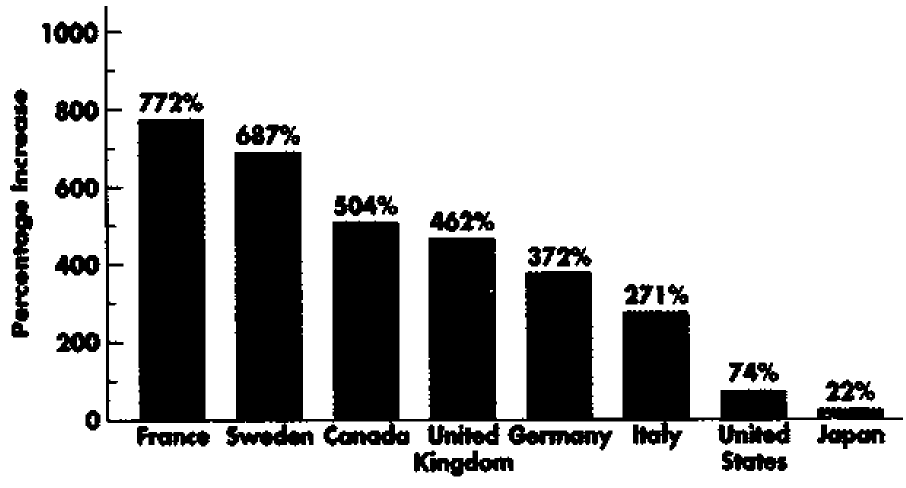
Long-term unemployment is a very different and more serious problem than brief intervals of joblessness. Four people, each unemployed for 3 months, will suffer lost income and temporary shocks. But the loss to one person unemployed for a year is likely much worse than the combined loss of the other four. Protracted unemployment is devastating psychologically, and is associated with increased psychosomatic illness, divorce, suicide, and other ills that are damaging to society as well as to the jobless worker and his or her family. Worse yet, there are particular groups who have reason to fear that they will never work again. This is true of workers aged 50 and over who have lost their jobs through corporate restructuring and downsizing and of young people in some rural and inner-city areas.

CESSATION OF WAGE GROWTH

By far the most dramatic recent development in the U.S. labor market is the one mentioned at the start of this chapter—the stagnation of real hourly earnings. Moreover, no one has any basis for a confident opinion about how long the problem is likely to continue, and we have only conjectures about its causes. Part of the problem has been attributed to the increased share of blacks, Hispanics, and women in the labor force—all traditionally low-wage-earning groups. Both the decline of union membership and increased foreign competition may have stim-

FIGURE 16-12

INCREASE IN LONG-TERM UNEMPLOYMENT AS A SHARE OF THE LABOR FORCE, 1975–1993



There has been a sharp increase in the share of the labor force that has been unemployed for 1 year or more, particularly in a number of European countries.

Note: For Japan and Canada, the graph reports data for 1979 to 1993; for the United Kingdom, Italy, and Germany, it reports data for 1975 to 1992.

SOURCES: Organization for Economic Cooperation and Development, *Employment Outlook, July 1994* (Paris: OECD, 1994); *The OECD Jobs Study: Facts, Analysis, Strategies* (Paris: OECD, 1994); *Quarterly Labour Force Statistics*, various issues; *Labour Force Statistics, 1971–1991* (Paris: OECD, 1993).

ulated downsizing that cut the number of positions offered by America's largest firms and added to downward pressure on wages.

The wage-decline problem is also often attributed to the slowing of the rate of growth in labor productivity that began in the late 1960s, since workers have usually been able to claim and get higher wages when their output per hour has gone up. However, this last point is not convincing. Even though productivity growth did *slow down*, its *level* did continue to rise. Nevertheless, real hourly earnings have failed to go up, and that is a matter for concern not only for workers, but for all Americans.

SUMMARY

1. The supply of labor is determined by free choices made by individuals. Because of conflicting *income and substitution effects*, the quantity of labor supplied may rise or fall as a result of an increase in wages.
2. Historical data show that hours of work per week have fallen as wages have risen, suggesting that income effects may be dominant.
3. The demand curve for labor, like the demand curve for any factor of production, is derived from the marginal revenue product curve. It slopes downward because of the "law" of diminishing marginal returns.
4. In a free market, the wage rate and the level of employment are determined by the interaction of supply and demand. Workers in great demand or short supply command high wages and, conversely, low wages go to workers in abundant supply or with skills not greatly demanded.
5. Some valuable skills are virtually impossible to duplicate. People who possess such skills will earn *economic rents* as part of their wages.
6. But most skills can be acquired by means of *investment in human capital*, such as education.
7. *Human capital theory* assumes that people make educational decisions in much the same way as businesses make investment decisions, and it tacitly assumes that people learn things in school that increase their productivity in jobs.
8. Other theories of the effects of education on earnings deny that schooling actually raises productivity. One view is that the educational system primarily sorts people according to their abilities. Another view holds that schools sort people according to their social class and teach them mainly discipline and obedience.
9. According to the theory of *dual labor markets*, there are two distinct types of labor markets, with very little mobility between them. The primary labor market contains the "good" jobs where wages are high, prospects for advancement are good, and higher education pays off. The secondary labor market contains the "bad" jobs with low wages, little opportunity for promotion, and little return to education.
10. About 15 percent of all American workers belong to *unions*, which can be thought of as monopoly sellers of labor. Compared with many other industrialized countries, unions in America are younger, less widespread, and less political.
11. For the most part, unions probably force wages to be higher and employment to be lower than they would be in a competitive labor market.
12. *Collective bargaining* agreements between labor and management are complex documents covering much more than employment and wage rates.
13. Strikes play an important role in collective bargaining as a way of dividing the fruits of economic activity between big business and big labor. Fortunately, strikes are not nearly so common as is often supposed.
14. For about two decades Americans have experienced three striking trends: a more than 30 percent decline in union membership, a rise in the proportion of the unemployed who are jobless for more than a year, and a steady fall in real wages partly offset by rising fringe benefits.

KEY TERMS

Minimum wage law
Income and substitution effects
Backward-bending supply curve
Economic rent
Investment in human capital
Human capital theory
Dual labor markets
Union
Industrial unions

Craft unions
Taft-Hartley Act (1947)
Closed shop
Union shop
Monopsony
Bilateral monopoly
Collective bargaining
Mediation
Arbitration

QUESTIONS FOR REVIEW

1. Colleges are known to pay rather low wages for student labor. Can this be explained by the operation of supply and demand in the local labor markets? Is the concept of monopsony of any use? How might things differ if students formed a union?
2. College professors are highly skilled (or at least highly educated!) labor. Yet their wages are not very high. Is this a refutation of the marginal productivity theory?
3. The following table shows the number of pizzas that can be produced by a large pizza parlor employing various numbers of pizza chefs.
   a. Find the marginal physical product schedule of chefs.
   b. Assuming a price of $5 per pizza, find the marginal revenue product schedule.
   c. If chefs are paid $70 per day, how many chefs will this pizza parlor employ? How would your answer change if chefs' wages rose to $95 per day?
   d. Suppose the price of pizza rises from $5 to $6. Show what happens to the derived demand curve for chefs.

Number of Chefs	Number of Pizzas per Day
1	40
2	64
3	82
4	92
5	100
6	92

4. Discuss the concept of the financial rate of return to a college education. If this return is less than the return on a bank account, does that mean you should quit college? Why might you wish to stay in school anyway? Are there circumstances under which it might be rational not to go to college, even when the financial returns to college are very high?
5. It seems to be a well-established fact that workers with more years of education typically receive higher wages. What are some possible reasons for this?
6. Approximately what fraction of the American labor force belongs to unions? (Try asking this question of a person who has never studied economics.) Why do you think this fraction is so low?
7. What are some reasonable goals for a union? Use the tools of supply and demand to explain how a union might pursue its goals, whatever they are. Consider a union that has been in the news recently. What was it trying to accomplish?
8. "Strikes are simply intolerable and should be outlawed." Comment.
9. In which of the following industries is wage determination most plausibly explained by the model of perfect competition? The model of pure monopoly? The model of bilateral monopoly? (a) Odd-job repairs in private homes; (b) Manufacture of low-priced clothing for children; (c) Steel manufacturing.
10. In a bitter strike battle between Eastern Airlines and several of its unions, it was clear from the beginning that the airline was in serious financial trouble. The airline was, indeed, eventually forced to close down at the cost of many jobs. Discuss what might nevertheless have led the unions to hold out so tenaciously.
11. Can you think of some types of workers whose marginal products probably were raised by computerization? Are there any whose marginal products were probably reduced? Can you characterize the difference between the two types of jobs in general terms?
12. European labor unions have traditionally had a strong socialistic orientation. How would you guess this is likely to be affected by the movement of countries in eastern Europe toward market economies?
13. What, if anything, do you think is the effect of long-term unemployment on crime rates? What about short-term unemployment?
14. Since about 1980 GDP per capita (that is, the average real income per person) in the United States has risen fairly substantially. Yet real wages have failed to rise. What do you think may explain this?

CHAPTER 17

POVERTY, INEQUALITY, AND DISCRIMINATION

The white man knows how to make everything, but he does not know how to distribute it.
Sitting Bull

The last two chapters analyzed how factor prices—wages, rents, and interest rates—are determined in a market economy. One reason for concern with this issue is that these payments determine the *incomes* of the people who own the factors. The study of factor pricing is, therefore, an indirect way to learn about the *distribution of income* among individuals.

In this chapter, we turn directly to the problem of income distribution. Specifically, we seek answers to the following questions: How much income inequality is there in the United States, and why? How can society decide rationally on how much equality it wants? And, once this decision is made, what policies are available to pursue this goal? In trying to answer these questions, we must necessarily consider the related problems of poverty and discrimination, and so these issues, too, are addressed in the chapter.

We will also offer a full explanation of one of the **Ideas for Beyond the Final Exam:** *the fundamental trade-off between economic equality and economic efficiency.* Taking it for granted that equality and efficiency are both important social goals, we shall learn why policies that promote greater income equality may interfere with economic efficiency. In this chapter we explain *why* this is so and *what* can be done about it.

ISSUE: WELFARE REFORM AND THE TRADE-OFF BETWEEN EQUALITY AND EFFICIENCY

For several decades, America's main programs for the poor have been widely decried as "the welfare mess"—a vague term indicating that they are too bureaucratic, too expensive, and too ineffective. Some critics have gone so far as to claim that welfare hurts the very people it is designed to help by, for example, encouraging out-of-wedlock births and fostering a culture of dependence on the state. When candidate Bill Clinton campaigned on a promise to "end welfare as we know it" in 1992, many Americans shared his desire.

But the Clinton administration quickly became embroiled in other issues, and nothing much happened until 1995. Then, however, radical changes in welfare were proposed. The 1994 elections had brought a far more conservative Congress to Washington—one which was determined not only to reform welfare, but to trim its costs and to give more authority to the states. Initially, President Clinton accepted some of these proposals. But he ultimately vetoed the Republican welfare bill in January 1996, leaving both welfare critics and advocates wondering what would happen next.

What is really at issue here? Are the criticisms of welfare justified? Or would criticism be more aptly directed at the proposed welfare reforms? As we shall see in this chapter, the debate over welfare reform is a classic example of the trade-off between equality and efficiency—an idea that is poorly understood by the general public. Some liberals argue that society should adopt even the most outlandish programs to reduce discrimination, increase income equality, or eradicate poverty—regardless of the potential side effects these policies might have. Some conservatives, on the other hand, seem so obsessed with these undesirable side effects—whether real or imagined—that they ignore the benefits of redistribution or of antidiscrimination programs.

Economists prefer to avoid absolutes and to think in terms of trade-offs: To reap gains on one front, you often must make sacrifices on another. A policy is not necessarily ill-conceived simply because it has an undesirable effect on income inequality, *if* it makes an important enough contribution to efficiency. But policies with very bad distributive consequences may deserve to be rejected, even if they would raise the GDP.

Admitting that there is a trade-off between equality and efficiency—that welfare spending may alleviate poverty but reduce economic efficiency—may not be the best way to win votes. But it does face the facts. And in that way it helps us make the inherently political decisions about what should be done. If we are to understand these complex issues, a good place to start is, as always, with the facts.

THE FACTS: POVERTY

In 1962, Michael Harrington published a little book called *The Other America*, which was to have a profound effect on American society. Harrington's "other Americans" were the poor who lived in the land of plenty. Ill-clothed in the richest country on earth, inadequately nourished in a nation where obesity was a problem, infirm in a country with some of the world's highest health standards, these people lived an almost unknown existence in their dilapidated hovels, according to Harrington. To make matters worse, this deprivation often condemned the children of the "other Americans" to repeat the lives of their parents. There was, Harrington argued, a "cycle of poverty" that could be broken only by government action.

The work of Harrington and others touched the hearts of many Americans who, it seemed, really had no idea of the abominable living conditions of some of their countrymen. Within a few years, the growing outrage over the plight of the poor had crystallized into a "War on Poverty," which was declared by President Lyndon Johnson in 1964.

The **POVERTY LINE** is an amount of income below which a family is considered "poor."

As part of this program, the government adopted an official definition of poverty: The poor were those families with incomes below $3,000 in 1964. This dividing line between the poor and nonpoor was called the **poverty line**, and a goal was established: to get all Americans above the poverty line by the nation's

bicentennial in 1976. (The goal was not met.) The poverty line was subsequently modified to account for differences in family size and other considerations, and it is now also adjusted each year to reflect changes in the cost of living. In 1994, the poverty line for a family of four was just over $15,000 and 14.5 percent of all Americans remained in poverty by official definitions.

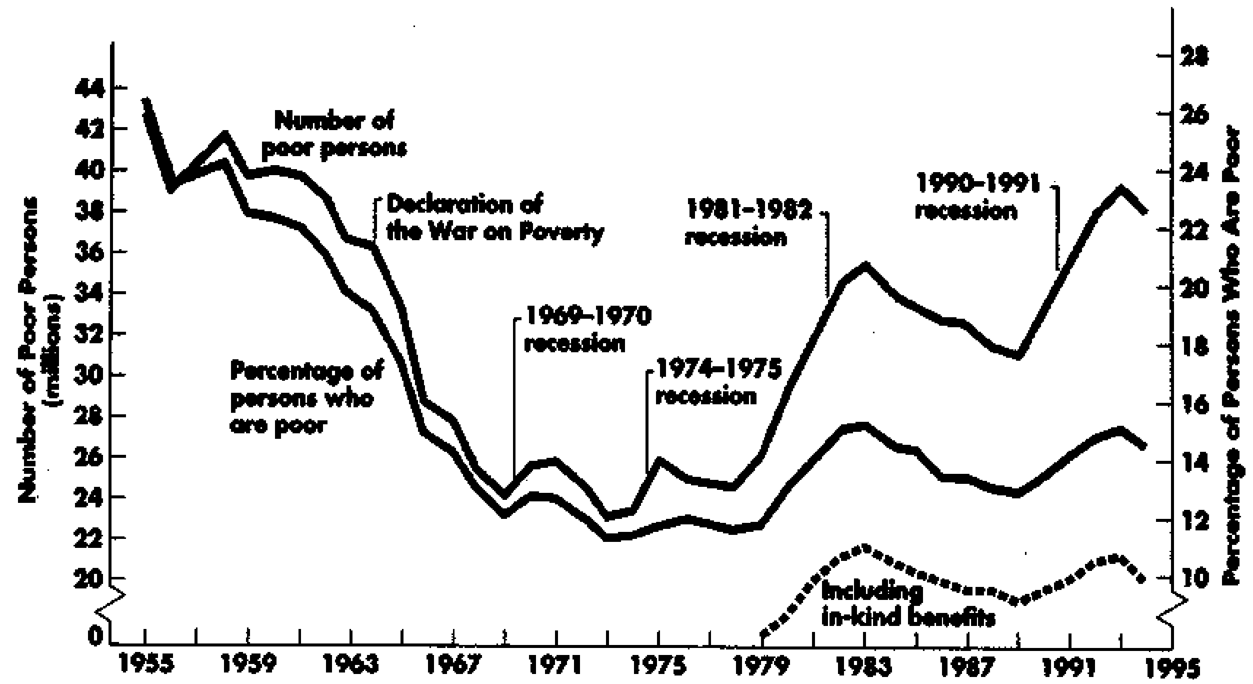
Who are the poor? Relative to their proportions in the overall population, they are more likely to be black than white and female than male. They are less educated and in worse health than the population as a whole. About 40 percent of the poor are children.

Substantial progress toward eliminating poverty was made in the decade from 1963 to 1973; the percentage of people living below the poverty line dropped from 20 percent to 11 percent (see the blue line in Figure 17-1). But thereafter slower economic growth and cutbacks in social welfare programs reversed the trend. By 1983, the poverty rate was back to what it had been in the 1960s. Since then, the poverty rate has increased and decreased with no clear trend, but it is still well above its 1970's low.

The rise in poverty since the 1970s worries many people, especially since poverty nowadays seems often to be associated with homelessness, illegitimacy, drug dependency, and ill health—all symptoms of a growing underclass whose lives are no better, and in many respects worse, than the people Harrington wrote about in 1962. (See the box on the next page.)

However, some critics argue that the official data badly overstate the number of poor persons. Some even go so far as to claim that poverty would be

FIGURE 17-1 PROGRESS IN THE WAR ON POVERTY



This figure charts the number and percentage of Americans classified as "poor" by official definitions. While substantial progress has been made in the War on Poverty, about 14.5 percent of Americans remain below the poverty line. The broken line shows one of the experimental measures of poverty that includes noncash benefits.

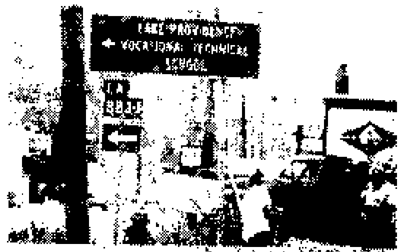
SOURCE: For 1959–1994, U.S. Bureau of the Census. For 1955–1958, estimates kindly provided by Gordon M. Fisher.

THE POOREST PLACE IN AMERICA

The town has no public parks or swimming pools, no movie theaters, no shopping malls, not even a McDonald's or a Wal-Mart. In fact, business in Lake Providence, Louisiana, is so bad that even the pawnshop has shut down. "The only recreation we have," says a resident, "is poor people's fun: drinking, drugs, fighting, and sex." Restless teenagers mill around narrow streets lined with burned-out houses and dilapidated trailer parks. . . . If there is a poorer place in America, the Census Bureau cannot find it. . . .

The 1990 census found that the median annual household income in Block Numbering Area 9903, which covers the southern two-thirds of Lake Providence and three-quarters of its population, was only $6,536—less than half the official poverty level of $14,764 for a family of four and the lowest in the United States. Two years later, a Children's Defense Fund study found that in East Carroll Parish, where Lake Providence is located, 70.1 percent of children younger than 18, or 2,409, were living in poverty, the highest rate in the nation—and this amid staggeringly high rates of infant mortality, teenage pregnancy, and drug use.

Meanwhile, jobs are scarce, low paying, and seasonal. For most of the year, hundreds of families subsist on welfare: A single mother with one child gets $123 a month. . . . For



many the only available work is backbreaking minimum-wage jobs in the nearby cotton fields. . . .

Inevitably, almost everyone who can escape from Lake Providence does so. "I'd rather shoot myself than stay here. It would be a wasted life," says Karva Henderson, who graduated from Lake Providence's high school in June. She plans to go to college and wants never to return. . . .

SOURCE: Jack E. White, "The Poorest Place in America," *Time*, August 15, 1994, pp. 35–36.

considered a thing of the past if the official definition (based on cash income) were amended to include the many goods that the poor are given in kind: public education, public housing, health care, food, and the like.

These criticisms prompted the Census Bureau to develop several experimental measures of poverty which include the value of goods given in kind. If these new measures are accepted as valid, fewer people are classified as poor, but the basic patterns of recent years are the same: Poverty rose sharply from 1979 to 1983, fell through 1988, and has risen since. (See the broken line in Figure 17-1.)

This debate raises a fundamental question: How do we define "the poor"? Continuing economic growth will eventually pull almost everyone above any arbitrarily established poverty line. Does this event mark the end of poverty? Some would say yes. But others would insist that the biblical injunction is right: "The poor ye have always with you."

There are two ways to define poverty. The more optimistic definition uses an *absolute concept of poverty:* If you fall short of a certain minimum standard of living, you are poor; once you pass this standard, you are no longer poor. The second definition is based on a *relative concept of poverty:* The poor are those who fall too far behind the average income.

Each definition has its pros and cons. The basic problem with the absolute poverty concept is that it is arbitrary. Who sets the line? Most of the people of Bangladesh would be delighted to live a bit below the U.S. poverty line and would consider themselves quite prosperous. Similarly, the standard of living that we now call "poor" would probably not have been considered so in America in 1780, and certainly not in Europe during the Middle Ages. Different times and different places apparently call for different poverty lines.

The fact that the concept of poverty is culturally, not physiologically, determined suggests that it must be a relative concept. For example, one suggestion

is to define the poverty line as one-half of the national average income. In this way, the poverty line would automatically rise as the nation grows richer.

Once we move from an absolute to a relative concept of poverty, the sharp distinction between the poor and the nonpoor starts to evaporate. Instead, we begin to think of a parade of people from the poorest soul to the richest billionaire. The "poverty problem" then becomes an issue of disparities in income that are "too large." At least in part, the poor are so poor because the rich are so rich. If we follow this line of thought far enough, we are led away from the narrow problem of *poverty* toward the broader problem of *inequality of incomes*.

THE FACTS: INEQUALITY

Nothing in the market mechanism guarantees income equality. On the contrary, the market tends to breed inequality, for the basic source of its great efficiency is its system of rewards and penalties. The market is generous to those who are successful in operating efficient enterprises that are responsive to consumer demands, and it is ruthless in penalizing those who are unable or unwilling to satisfy consumer demands efficiently.

Its financial punishment of those who try and fail can be particularly severe. At times it even brings down the great and powerful. Robert Morris, once perhaps the wealthiest resident of the American colonies, ended up in debtors' prison. In more recent decades, the financial travails of the Hunt brothers of Texas and Donald Trump of New York, once among America's richest people, have been highly publicized.

Most people have a pretty good idea that the gulf between the rich and the poor is wide. But few have any concept of where they stand in the income distribution. For example, during the 1995 congressional debate over cutting taxes for "the middle class," one member of Congress with an annual income in excess of $150,000 declared himself a member of the "middle class," if not indeed of the "lower-middle class"!

Table 17-1 offers some statistics on the 1994 income distribution in the United States. But before looking at them, try the following experiment. First, write down what you think your family's before-tax income was in 1994. (If you do not know, take a guess.) Next, try to guess what percentage of American families had incomes *lower* than this. Finally, if we divide America into three broad income classes—rich, middle class, and poor—to which group do you think your family belongs?

Now that you have written down answers to these three questions, look at the income distribution data for 1994 in Table 17-1. If you are like most college students, these figures may surprise you. First, if we adopt the tentative definition that the lowest 20 percent are the "poor," the highest 20 percent are the "rich," and the middle 60 percent are the "middle class," many fewer of you belong to the celebrated "middle class" than thought they did. In fact, the cutoff point that defined membership in the "rich" class in 1994 was only about $70,000 before taxes, an income level exceeded by the parents of many college students. (Your parents may be shocked to learn that they are rich!)

Next, use Table 17-1 to estimate the fraction of U.S. families that have incomes lower than your family's. (The caption has instructions to help you do this.) Most students who come from households of moderate prosperity have an instinctive feeling that they stand somewhere near the middle of the income distribution;

TABLE 17-1

DISTRIBUTION OF FAMILY INCOME IN THE UNITED STATES IN 1994

Income Range (dollars)	Percentage of All Families in This Range	Percentage of Families in This and Lower Ranges
Under 5,000	3.1	3.1
5,000 to 9,999	5.6	8.7
10,000 to 14,999	6.9	15.6
15,000 to 24,999	15.0	30.6
25,000 to 34,999	14.3	44.9
35,000 to 49,999	18.0	62.9
50,000 to 74,999	19.9	82.8
75,000 to 99,999	8.8	91.6
100,000 and over	8.4	100.0

If your family's income falls close to one of the end points of the ranges indicated here, you can approximate the fraction of families with income lower than yours by just looking at the last column.

If your family's income falls within one of the ranges, you can interpolate the answer. Example: Your family's income was $45,000. This is two-thirds of the way from $35,000 to $50,000, so your family was richer than roughly ($\frac{2}{3}$) × 18.0 percent = 12.0 percent of the families in this class. Adding this to the percentage of families in lower classes (44.9 percent in this case) gives the answer—about 56.9 percent of all familes earned less than yours.

SOURCE: U.S. Bureau of the Census.

so they estimate about half, or perhaps a little more. In fact, the median income among American families in 1994 was only about $39,000.

This exercise has perhaps brought us down to earth. America is not nearly as rich as Madison Avenue would like us to believe. Let us now look past the average level of income and see how the pie is divided. Table 17-2 shows the shares of income accruing to each fifth of the population in 1994 and several earlier years. In a perfectly equal society, all the numbers in this table would be "20 percent" since each fifth of the population would receive one-fifth of the income. In fact, as the table shows, this is far from true. In 1994, for example, the poorest fifth of all families had just over 4 percent of the total income, while the richest fifth had nearly 47 percent—about 11 times as much.

TABLE 17-2 INCOME SHARES IN SELECTED YEARS

Income Group	1994	1990	1980	1970	1960	1950
Lowest fifth	4.2	4.6	5.1	5.4	4.8	4.5
Second fifth	10.0	10.8	11.6	12.2	12.2	12.0
Middle fifth	15.7	16.6	17.5	17.6	17.8	17.4
Fourth fifth	23.3	23.8	24.3	23.8	24.0	23.4
Highest fifth	46.9	44.3	41.6	40.9	41.3	42.7

SOURCE: U.S. Bureau of the Census.

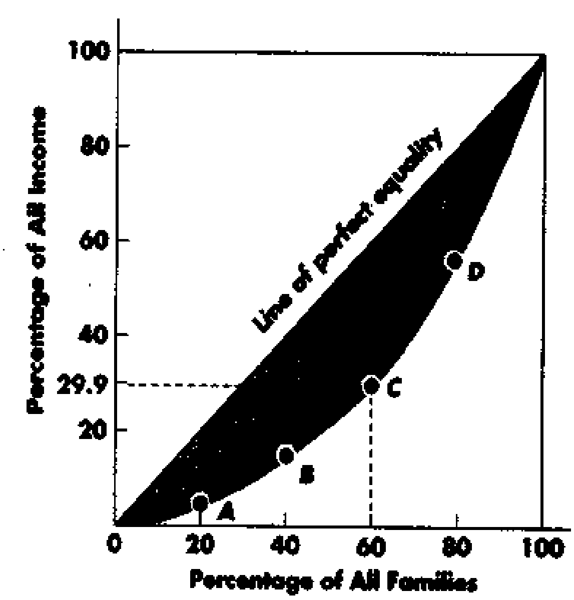
DEPICTING INCOME DISTRIBUTIONS: THE LORENZ CURVE

Statisticians and economists use a convenient tool to portray data like these graphically. The device, called a *Lorenz curve,* is shown in Figure 17-2. To construct a Lorenz curve, we first draw a square whose vertical and horizontal dimensions both represent 100 percent. Then we record the percentage of families (or persons) on the horizontal axis and the percentage of income that these families (or persons) receive on the vertical axis, using all the data that we have. For example, point C in Figure 17-2 depicts the fact (known from Table 17-2) that in 1994 the bottom 60 percent (the three lowest fifths) of American families received 29.9 percent of the total income. Similarly, points *A, B,* and *D* represent the other information contained in Table 17-2. We can list four important properties of a Lorenz curve.

1. It begins at the origin because zero families naturally have zero income.
2. It always ends at the upper-right corner of the square, since 100 percent of the nation's families must receive all the nation's income.
3. If income were distributed equally, the Lorenz curve would be a straight line connecting these two points (the thin, solid line in Figure 17-2). This is because, with everybody equal, the bottom 20 percent of the families would receive 20 percent of the income, the bottom 40 percent would receive 40 percent, and so on.
4. In a real economy, with significant income differences, the Lorenz curve will "sag" downward from this line of perfect equality. It is easy to see why this is so. If there is any inequality at all, the poorest 20 percent of families must get less than 20 percent of the income. This corresponds to a point below the equality line, such as point *A*. Similarly, the bottom 40 percent of families must receive less than 40 percent of the income (point *B*), and so on.

FIGURE 17-2 A LORENZ CURVE FOR THE UNITED STATES

This Lorenz curve for the United States is based on the 1994 distribution of income given in Table 17-2. The percentage of families is measured along the horizontal axis, and the percentage of income that these families receive is measured along the vertical axis. Thus, for example, point C indicates that the bottom 60 percent of American families received 29.9 percent of the total income in 1994.



In fact, the size of the area between the line of perfect equality and the Lorenz curve (the shaded area in Figure 17-2) is often used as a handy measure of inequality. The larger this area, the more unequal is the income distribution. For U.S. family incomes, this so-called *area of inequality* fills about 40 percent of the total area underneath the equality line.

By itself, the Lorenz curve tells us little. To interpret it, we must know what it looked like in earlier years or what it looks like in other countries.

The historical data in Table 17-2 show that *the U.S. Lorenz curve has not moved much since World War II.* To some, this remarkable stability in the income distribution is deplorable. To others, it suggests some immutable law of the capitalist system. Notice, however, that the distribution of income grew more equal in the early part of the postwar period, but has grown substantially more unequal since. The share of the poorest fifth is now the lowest, and the share of the richest fifth is now the highest since the government began collecting data in 1947.

The distribution of income in the United States grew slightly more equal from the 1950s to the 1970s, but has grown more unequal during the 1980s and 1990s.

America is not a very class-conscious society, and for years only specialists paid much attention to data like those in Table 17-2. But income inequality has recently become a big social issue, as more and more American families sense that they are losing ground to the people at the top. There is particular, and well-justified, concern that the real earnings of wage-earners below the middle have fallen dramatically in the past two decades.

Comparing the United States with other countries is much harder, since no two countries use precisely the same definition of income distribution. In 1995, the Organization for Economic Cooperation and Development (OECD) published the first study in almost 20 years to use standardized data to compare the income distributions of different countries. Among the 16 (mostly European) countries included in the analysis, Sweden and Finland had the most equal income distributions, while the United States and Ireland had the most unequal. Thus, it appears that:

The United States has rather more income inequality than most other industrialized countries.

It also appears that the rise in income inequality in the United States since the 1970s is unusually sharp by world standards.

SOME REASONS FOR UNEQUAL INCOMES

Let us now begin to formulate a list of the *causes* of income inequality. Here are some that come to mind.

1. *Differences in ability.* Everyone knows that people have different capabilities. Some can run faster, ski better, do calculations more quickly, type more accurately, and so on. Hence, it should not be surprising that some people are more adept at earning income. Precisely what sort of ability is relevant to earning income is a matter of intense debate among economists, sociologists, and psychologists. The talents that make for success in school seem to have some effect, but hardly an overwhelming one. The same is true of innate intelligence—"IQ" (see the box on the next page). It is clear that some types of inventiveness are richly rewarded by the market, as is that elusive charac-

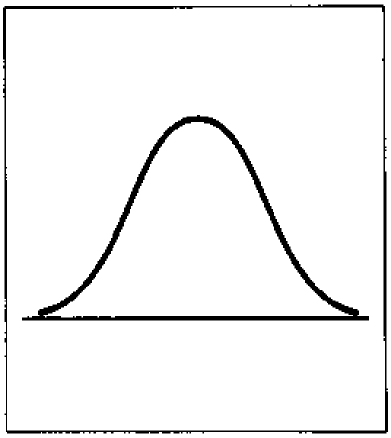
HOW IMPORTANT IS THE BELL CURVE?

In 1994, social critic Charles Murray and the late Richard Herrnstein, a psychologist, created a furor with a book claiming that genetically inherited intelligence is an overwhelmingly important determinant of economic success. The book's title, *The Bell Curve*, was a reference to the shape of the distribution of observed test scores on conventional IQ tests (see chart), which show most people clustered near the middle of the distribution, with small minorities on either end.

Critics of government antipoverty programs were quickly attracted to the book's central message: that the poor are poor in large measure because they are not very smart. Among the most stunning claims made by Herrnstein and Murray was that much of the observed economic gap between blacks and whites could be attributed to the fact that blacks' IQs were, on average, lower than those of whites.

While *The Bell Curve* received a blitz of media attention, social scientists generally ignored it or gave the analysis low marks. No one seriously doubts that intelligence contributes to economic success, nor that genetics has some bearing on intelligence. But the scientific evidence on the strength of each link is in great dispute. Many experts on IQ, for example, argue that environmental factors may be more important than genetics in determining intelligence—and that "true" intelligence may differ from measured IQ. Furthermore, few if any economists believe that cognitive ability is the main ingredient in economic success.

The bottom line, according to most scholars, is that the black-white IQ gap does not go very far in explaining racial income inequalities. Nor can we be certain that much of the measured IQ gap is biologically, rather than culturally, determined.



teristic called "entrepreneurial ability." Also, it is obvious that poor health often impairs earning ability.

2. *Differences in intensity of work.* Some people work longer hours than others, or labor more intensely when they are on the job, leading to income differences that are largely voluntary.
3. *Risk taking.* Most people who acquire large sums of money do so by taking risks—by investing their money in some uncertain venture. Those who gamble and succeed become wealthy. Those who try and fail go broke. Most others prefer not to take such chances and end up somewhere in between. This is another way in which income differences arise voluntarily.
4. *Compensating wage differentials.* Some jobs are more arduous than others, or more dangerous, or more unpleasant for other reasons. To induce people to take these jobs, some sort of financial incentive normally must be offered. For example, factory workers who work the night shift normally receive higher wages than those who work during the day.
5. *Schooling and other types of training.* Chapter 16 analyzed schooling and other types of training as "investments in human capital." The term refers to the idea that workers can sacrifice *current* income in order to improve their skills so that their *future* incomes will be higher. When this is done, income differentials naturally rise.

While it is generally agreed that differences in schooling are an important cause of income differentials, this particular cause has both voluntary and involuntary aspects. Young men or women who *choose* not to go to college

have made voluntary decisions that affect their incomes. But many never get the choice: Their parents simply cannot afford to send them. For them, the resulting income differential is not voluntary.

6. *Work experience.* It is well-known to most people and well-documented by scholarly research that more experienced workers earn higher wages.
7. *Inherited wealth.* Not all income is derived from work. Some is the return on invested wealth, and part of this wealth is inherited. While this cause of inequality applies to few people, many of America's super-rich got that way through inheritance.

   And financial wealth is not the only type of capital that can be inherited; so can human capital. In part this happens naturally through genetics: High-ability parents tend to have high-ability children, although the link is an imperfect one. But it also happens partly for economic reasons: Well-to-do parents send their children to the best schools, thereby transforming their own *financial* wealth into *human* wealth for their children. This type of inheritance may be much more important than the financial type.
8. *Luck.* No observer of our society can fail to notice the role of chance. Some of the rich and some of the poor got there largely by good or bad fortune. A farmer digs for water, but strikes oil instead. A student prepares diligently for a high-paying occupation only to find that the opportunity has disappeared before graduation. A construction worker is unemployed for a whole year because of a recession that he had no part in creating. The list could go on and on. Many large income differentials arise purely by chance.

THE FACTS: DISCRIMINATION

Some of the factors we have just listed lead to income differentials that are widely accepted as "just." For example, most people believe it is fair for people who work harder to receive higher incomes. Other factors on our list ignite heated debates. For example, some people view income differentials that arise purely by chance as perfectly acceptable. Others find these same differentials intolerable. However, almost no one is willing to condone income inequalities that arise from discrimination.

ECONOMIC DISCRIMINATION occurs when equivalent factors of production receive different payments for equal contributions to output.

The facts about discrimination are not easy to come by. **Economic discrimination** is defined to occur when equivalent factors of production receive different payments for equal contributions to output. But this definition is hard to apply in practice because we cannot always tell when two factors of production are "equivalent."

If a woman with only a high school diploma receives a lower salary than a man with a college degree, few people would call that "discrimination." Even if a man and a woman have the same education, the man may have 10 more years of work experience than the woman. If they receive different wages for this reason, is that discriminatory?

In principle, we should compare men and women whose *productivities* are equal. If women receive lower wages than men, we would then attribute the difference to discrimination. But discrimination normally takes much more subtle forms than paying unequal wages for equal work. For instance, employers can simply relegate women to inferior jobs, thus justifying lower salaries.

One clearly *incorrect* way to measure discrimination is to compare the typical incomes of different groups. Table 17-3 displays such data for white men, white

TABLE 17-3 MEDIAN INCOMES IN 1994

Population Group[a]	Median Income	Percentage of White, Male Income
White males	$22,669	100
Black males	14,982	66
White females	11,630	51
Black females	10,544	47

[a]Persons 15 years old and over.
SOURCE: U.S. Bureau of the Census.

women, black men, and black women in 1994. Virtually everyone agrees that the amount of discrimination is less than these differentials suggest, but far greater than zero. Precisely how much is a topic of continuing economic research. Several studies suggest that about half of the observed wage differential between black and white men, and at least half of the differential between white women and white men, is caused by discrimination in the labor market (though more might be due to discrimination in education, and so on). Other studies have reached somewhat different conclusions. While no one denies the existence of discrimination, its quantitative importance is a matter of ongoing controversy and research.

THE ECONOMIC THEORY OF DISCRIMINATION[1]

Let us see what economic theory tells us about discrimination. In particular, consider the following two questions:

1. Must *prejudice*, which we define as arising when one group dislikes associating with another group, lead to *discrimination* (unequal pay for equal work)?
2. Are there "natural" economic forces that tend either to erode or to exacerbate discrimination over time?

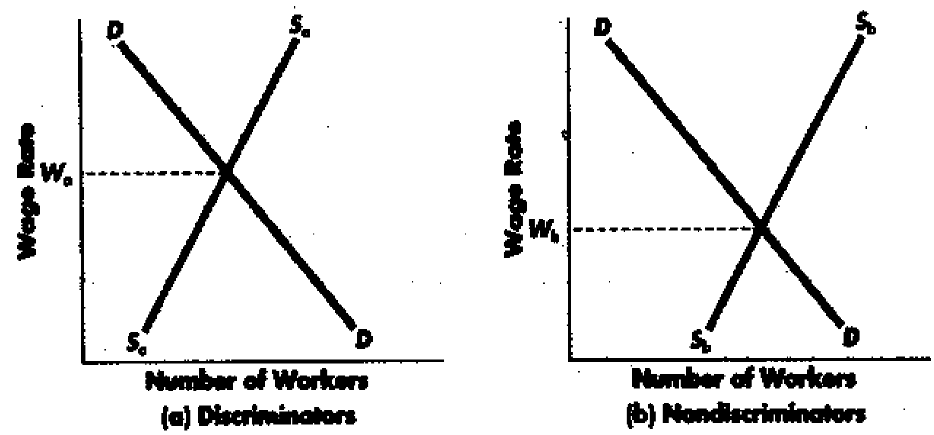
As we shall see now, the analysis we have provided in previous chapters sheds light on both these issues.

DISCRIMINATION BY EMPLOYERS

Most attention seems to focus on discrimination by employers, so let us start there. What happens if, for example, some firms refuse to hire blacks? Figure 17-3 will help us find the answer. Panel (a) pertains to firms that discriminate; Panel (b) pertains to firms that do not. There are supply and demand curves for labor in each market, based on the analysis of Chapter 16. We suppose the two demand curves to be identical. However, the supply curve in Market (b) must be farther to the right than the supply curve in Market (a) because whites *and* blacks can work in Market (b) whereas only whites can work in Market (a). The result is that wages will be lower in Market (b) than in Market (a). Since all the

[1]This section may be omitted in shorter courses.

FIGURE 17-3 WAGE DISCRIMINATION



Panel (a) depicts supply and demand curves for labor among discriminatory firms; Panel (b) shows the same for nondiscriminatory firms. Since only whites can work in Market (a), while both races can work in Market (b), the supply curve in Market (b) is farther to the right than the supply curve in Market (a). Consequently, the wage rate in Market (b), W_b, winds up below the wage rate in Market (a), W_a. Workers in Market (b) are discriminated against.

blacks are forced into Market (b), we conclude that they are discriminated against.

But now consider the situation from the point of view of *employers.* Firms in Market (a) of Figure 17-3 are paying more for labor, so they are paying for the privilege of discriminating. The nondiscriminatory firms in Market (b) have a cost advantage. As we learned in earlier chapters, if there is effective competition, these nondiscriminatory firms will tend to capture more and more customers. The discriminators will gradually be driven out of business. If, on the other hand, many of the firms in Market (a) have protected monopolies, they will be able to remain in business. But they will pay for the privilege of discriminating by earning lower monopoly profits than they otherwise could (because they pay higher wages than they have to pay).

DISCRIMINATION BY FELLOW WORKERS

Thus, competitive forces will tend to reduce discrimination over time *if* employers are the source of discrimination. Such optimistic conclusions cannot necessarily be reached, however, if it is workers who are prejudiced. Consider what happens if, for example, men do not like to have women as their supervisors. If male workers do not give their full cooperation, female supervisors will be less effective than male supervisors and hence will earn lower wages. Here prejudice does lead to discrimination. Furthermore, in this case, firms that put women into supervisory positions will be at a competitive disadvantage relative to firms that do not. So market forces will not erode discrimination.

STATISTICAL DISCRIMINATION

STATISTICAL DISCRIMINATION is said to occur when the productivity of a particular worker is estimated to be low just because that worker belongs to a particular group (such as women).

A final type of discrimination, called **statistical discrimination,** may be the most stubborn of all and can exist even when there is no prejudice. Here is an important example. It is, of course, a fact that only women can have babies. It is also a fact that most working women who have babies leave their jobs for a while to care for their newborns. Employers know this. What they cannot know, however, is *which* women of child-bearing age are likely to leave the labor force for this reason.

Suppose three candidates apply for a job that requires a long-term commitment. Susan plans to quit after a few years to raise a family. Jane does not plan

ARE WOMEN BETTER WORKERS?

Economist Audrey Freedman argues that female employees can be a better bargain than male employees, even though only women request pregnancy leaves and it is mainly women who miss workdays for child-care reasons.

> It is undeniable ... that women, not men, take pregnancy leaves. It is also undeniable that women are the primary nurturers in a family. They are the most likely to be responsible for the care and support of children, as well as their elderly parents. If we stop there, ... women in business are more costly than men.
>
> But the built-in bias of that analysis is the failure to account for far more costly drains on corporate productivity from behavior that is more characteristic of men than of women.
>
> For example, men are more likely to be heavy users of alcohol.... This gender-related habit causes businesses to suffer excessive medical costs, serious performance losses, and productivity drains. Yet the male-dominated corporate hierarchy most often chooses to ignore these "good old boy" habits....
>
> Drug abuse among the fast-movers of Wall Street seems to be understood as a normal response to the pressures of taking risks with other people's money. The consequences in loss of judgment are tolerated. They are not calculated as a male-related cost of business.
>
> Apart from performance problems at high levels, alcohol and drug abuse causes costly accidents. We never think of them, however, as a risk primarily associated with male employees....
>
> In our culture, lawlessness and violence are found far more often among men than women. The statistics on criminals and prison population are obvious; yet we seem to be unable to recognize this as primarily male behavior....
>
> A top executive of a major airline once commented to me that his company's greatest problem is machismo in the cockpit—pilots and copilots fighting over the controls. There is an obvious solution: Hire pilots from that half of the population that is less susceptible to the attacks of rage that afflict macho males.

SOURCE: Audrey Freedman, "Those Costly 'Good Old Boys,'" *The New York Times*, July 12, 1989, p. A23.



to have any children. Jack is a man. If he knew all the facts, the employer might not want Susan but be indifferent between Jane and Jack. But the employer cannot tell Susan and Jane apart. He therefore presumes that both Jane and Susan, being young women, are more likely than Jack to quit to raise a family. So he hires Jack, even though Jane is just as good a prospect. Jane is discriminated against.

Lest it be thought that this example actually justifies discrimination against women on economic grounds, it should be noted that most women return to work within 6 months after childbirth. Furthermore, women typically have less absenteeism and job turnover for nonpregnancy health reasons than men do. The accompanying box argues that employers often fail to take these other sex-related differences into account, and thus mistakenly favor men.

THE ROLES OF THE MARKET AND THE GOVERNMENT

In terms of the two questions with which we began this section, we conclude that different types of *discrimination* lead to different answers. Prejudice often, but not always, leads to economic discrimination. And discrimination may occur even in the absence of prejudice. Finally, the forces of competition tend to erode some, but not all, of the inequities caused by discrimination.

However, the victims of discrimination are not the only losers. Society also loses whenever discriminatory practices impair economic efficiency. Hence, most observers believe that we should not rely on market forces *alone* to combat discrimination. The government has a clear role to play.

THE OPTIMAL AMOUNT OF INEQUALITY

We have seen that substantial income inequality exists in America, and we have noted some reasons for it. Let us now ask a question that is loaded with value judgments, but to which economic analysis has something to contribute nonetheless: *How much inequality is the ideal amount?* We shall not, of course, be able to give a definitive answer to this question. Rather, our objective is to see the type of analysis that is relevant. We begin in a simple setting in which the answer is easily obtained. Then we shall see how the real world differs from this simple model.

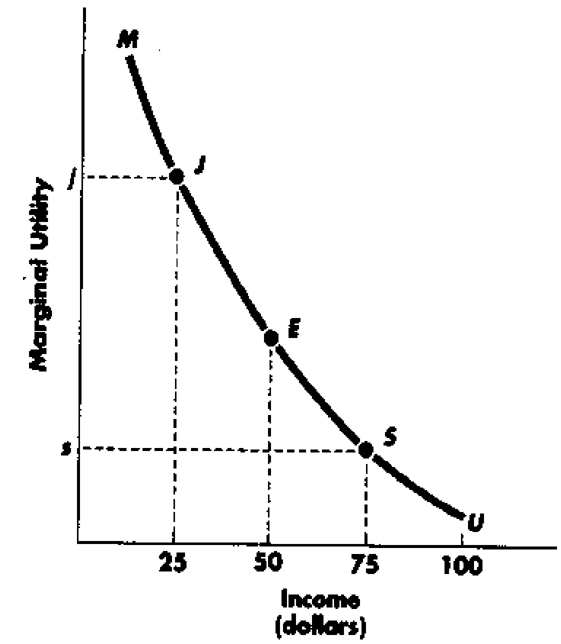
Consider a society in which two people, Smith and Jones, are to divide $100 between them. The objective is to maximize *total utility*. Suppose Smith and Jones are alike in their ability to enjoy money; technically, we say that their *marginal utility* schedules are identical.[2] This identical marginal utility schedule is depicted in Figure 17-4. We can prove the following result: *The optimal distribution of income is to give $50 to Smith and $50 to Jones*, which is point *E* in Figure 17-4.

We prove it by showing that, if the income distribution is unequal, we can improve things by moving closer to equality. So suppose that Smith has $75 (point *S* in the figure) and Jones has $25 (point *J*). Then, as we can see, Smith's *marginal utility* (which is *s*) must be *less* than Jones's (which is *j*). This is a simple consequence of the law of diminishing marginal utility.

If we take $1 away from Smith, he *loses* the low marginal utility, *s*, of a dollar to him. Then, when we give it to Jones, he *gains* the high marginal utility, *j*,

FIGURE 17-4 THE OPTIMAL DISTRIBUTION OF INCOME

If Smith and Jones have the identical marginal utility schedule (curve *MU*), then the optimal way to distribute $100 between them is to give $50 to each (point *E*). If income is not distributed this way, then their marginal utilities will be unequal, so that a redistribution of income can make society better off. This is illustrated by points *J* and *S*, representing an income distribution in which Jones gets $25 (and hence has marginal utility *j*) while Smith gets $75 (and hence has marginal utility *s*).



[2]If you need to refresh your memory about marginal utility, see Chapter 5, especially pages 97–100.

that a dollar gives him. On balance, society's total utility rises by $j - s$ because Jones's gain exceeds Smith's loss. Therefore, a distribution with Smith getting only $74 is better than one in which he gets $75. Since we can use the same argument to show that a $73/$27 distribution is better than $74/$26, and so on, we have established our result that a $50/$50 distribution—point *E*—is best.

Now in this example there is nothing special about the fact that we assumed only two people or that exactly $100 was available. Any number of people and dollars would do as well. What really *is* crucial is our assumption that the same amount of money would be available no matter how we chose to distribute it. Thus, we have proved the following general result:

To maximize total utility, the best way to distribute any *fixed* amount of money among people with identical marginal utility schedules is to divide it equally.

THE TRADE-OFF BETWEEN EQUALITY AND EFFICIENCY

If we seek to apply this analysis to the real world, two major difficulties arise. First, people are different and have different marginal utility schedules. Thus, *some* inequality can probably be justified.[3] The second problem is much more formidable.

The total amount of income in society is *not* independent of how we try to distribute it.

To see why, consider an extreme example: Ask yourself what would happen if we tried to achieve perfect equality by putting a 100 percent income tax on all workers and then dividing the tax receipts equally among the population. No one would have any incentive to work, to invest, to take risks, or to do anything else to earn money, because the rewards for all such activities would disappear. The gross domestic product (GDP) would fall drastically. While the example is extreme, the principle is universal; indeed, it is the basic idea behind supply-side economics.

THE TRADE-OFF BETWEEN EQUALITY AND EFFICIENCY

Policies that redistribute income reduce the rewards of high-income earners while raising the rewards of low-income earners. Hence, they reduce the incentive to earn high income. This gives rise to a trade-off that is one of the most fundamental in all of economics, and one of our **Ideas for Beyond the Final Exam.**

Measures taken to increase the amount of economic equality will often reduce economic efficiency—that is, lower the gross domestic product. In trying to divide the pie more equally, we may inadvertently reduce its size.

Because of this trade-off, equal incomes are not optimal in practice. On the contrary:

The optimal distribution of income will always involve *some* inequality.

[3]It can be shown that if we know that people differ, but cannot tell who has the higher marginal utility schedule, then the best way to distribute income is still in equal shares.

But this does not mean that attempts to reduce inequality are misguided. We should learn two things from this analysis:

1. There are better and worse ways to promote equality. In pursuing further income equality (or fighting poverty), we should seek policies that do the least possible harm to incentives.
2. Equality is bought at a price. Thus, like any commodity, we must decide rationally how much to purchase. We will probably want to spend some of our potential income on equality, but not all of it.

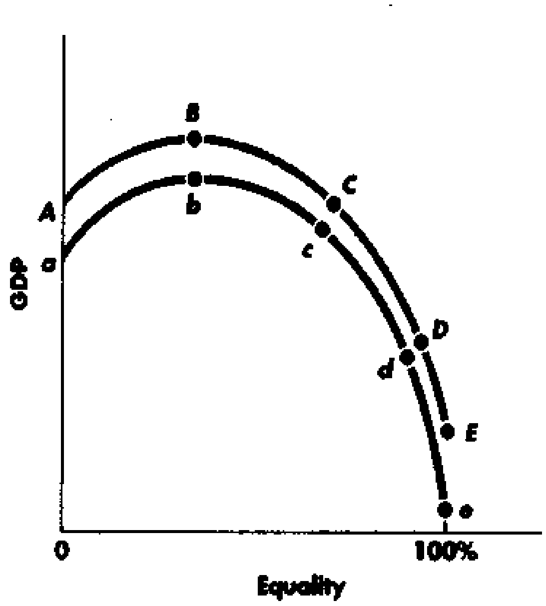
Figure 17-5 illustrates both of these lessons. The curve *abcde* represents possible combinations of GDP and income equality that are obtainable under the present system of taxes and transfers. If, for example, point *c* is the current position of the economy, raising taxes on the rich to finance more transfers to the poor, as President Clinton did in 1993, might move us downward to the right toward point *d*. Equality increases, but GDP falls as the rich react to higher marginal tax rates by producing less. Similarly, reducing both taxes and social welfare programs, as the Republican Congress proposed to do in 1995, might move us upward to the left toward point *b*. Notice that, to the left of point *b*, GDP falls as inequality rises. Here there is no trade-off—perhaps because very poorly paid workers are less productive due to inadequate investment in human capital, poor nutrition, or just a general sense of disaffection.

The curve *ABCDE* represents possible combinations of GDP and equality under some new, more efficient redistributive policy. It is more efficient in the sense that, for any desired level of equality, we can get more GDP with the policy represented by *ABCDE* than with the policy represented by *abcde*.

The first lesson is obvious: We should stick to the higher of the two curves. If we find ourselves at any point on curve *abcde*, we can always improve things by moving up to the corresponding point on curve *ABCDE*, that is, by changing policies. By picking the most efficient redistributive policy, we can have more equality *and* more GDP. In the rest of this chapter, we discuss alternative policies and try to indicate which ones do the least harm to incentives.

FIGURE 17-5 THE TRADE-OFF BETWEEN EQUALITY AND EFFICIENCY

This diagram portrays the fundamental trade-off between equality and efficiency. If the economy is initially at point *c*, then movements toward greater equality (to the right) normally can be achieved only by reducing economic efficiency, and thus reducing the gross domestic product. The movements from points *C* and *c* toward points *D* and *d* represent two alternative policies for equalizing the income distribution. The policy that leads to *D* is preferred since it is more efficient.



The second lesson is that neither point *B* nor point *E* would normally be society's optimal choice. At point *B* we are seeking the highest possible GDP with utter disregard for whatever inequality might accompany it. At point *E* we are forcing complete equality, even if work incentives vanish and a minuscule GDP is the result.

It is astonishing how much confusion is caused by a failure to understand these two lessons. Proponents of greater equality often feel obliged to deny that the programs they advocate will do any harm to incentives. Sometimes these vehement denials are so patently unrealistic that they undermine the very case that the egalitarians are trying to make. Conservatives who oppose such policies also undercut the strength of their case by making outlandish claims about the efficiency losses from redistribution.

Neither side, it seems, is willing to acknowledge the fundamental trade-off between equality and efficiency depicted in Figure 17-5. And so the debate generates more heat than light. Since these debates are likely to continue for the rest of your lives, we hope that some understanding of this trade-off stays with you well **Beyond the Final Exam.**

But merely understanding the trade-off will not tell you what to do. By looking at Figure 17-5, we know that the optimal amount of equality lies between points *B* and *E*, but we do not know what it actually is. Is it more like point *D*, with greater equality and less GDP than we now have? Or is it more like a movement back toward point *B*? Everyone will have a different answer to this question, because it is basically one of value judgments. Just how much is more equality worth to you?

Arthur Okun, once chairman of the Council of Economic Advisers, put the issue graphically. Imagine that money is liquid, and that you have a bucket to use to transport money from the rich to the poor. But the bucket leaks. As you move the money, some gets lost. Will you use the bucket if only 1 cent is lost for each $1 you move? Probably everyone would say yes. But what if each $1 taken from the rich results in only 10 cents for the poor? Only the most extreme egalitarians will still say yes. Now try the hard questions. What if 20 to 40 cents is lost for each $1 that you move? If you can answer questions like these, you can decide how far down the hill from point *B* you think society should travel, for you will have expressed your value judgments in quantitative terms.

POLICIES TO COMBAT POVERTY

Let us take it for granted that the nation has a commitment to reduce poverty. What are some policies that can promote this goal? Which of these does the least harm to incentives, and hence is most efficient?

EDUCATION AS A WAY OUT

Education is often thought of as one of the principal ways to escape from poverty. There is no doubt that many people have used this route successfully, and still do.[4] However, delivering quality education to the children of the poor is no simple matter. Many of them, especially in the inner cities, are ill-equipped to learn and attend schools that are ill-equipped to teach. Dropout rates are staggering. An astonishing number of youths leave the public school system without

[4] The role of education as a determinant of income was considered at length in the previous chapter.

even acquiring basic literacy. All of these problems are familiar; none is easy to solve.

In truth, our educational system is designed to serve many goals; and the alleviation of poverty is not the major one. If it were, we would almost certainly spend more money on, say, preschool and remedial education and less on college education. Furthermore, education is not a particularly effective way to lift *adults* out of poverty. Its effects take a generation or more to be realized.

THE WELFARE DEBATE

By contrast, a variety of programs collectively known as "welfare" are specifically designed to alleviate poverty, are meant to help adults as well as children, and are intended to have quick effects. The best known, and most controversial, of these is *Aid to Families with Dependent Children (AFDC)*. This program provides direct cash grants to families in which there are children but no breadwinner, perhaps because there is no father and the children are too young to permit the mother to work. In 1993, about 14 million people received benefits from AFDC, and the average monthly grant was about $373 per family. In total, the government spent more than $22 billion.

AFDC has been attacked as a classic example of an inefficient redistributive program, and was a major target of congressional budget-cutters in 1995 and 1996. Why? One reason is that it provides no incentive for the mother to earn income. Once monthly earnings pass a few hundred dollars, welfare payments are reduced by $1 for each $1 that the family earns as wages. Thus, if a member of the family gets a job, the family is subjected to a 100 percent marginal tax rate. It is little wonder that many welfare recipients do not look very hard for work.

A second criticism is that AFDC undermines traditional family structures by providing financial incentives for families to break up and for out-of-wedlock births. Originally, welfare benefits were not paid to a family with a father who could work, even if he was unemployed. So, if this father earned little or nothing, the children would get more income if he left home. Some fathers did. All the states reacted to this problem by amending their AFDC programs to pay benefits to families with unemployed fathers.

But the charge that welfare encourages childbearing by unwed mothers is still very much alive and, in fact, played a prominent role in the welfare debate in 1995 and 1996. Critics in Congress and elsewhere argue that AFDC encourages poor, unwed mothers to have more children because each additional child brings higher welfare payments. Their suggested solution? Stop increasing benefits after, say, two children. Defenders of welfare counter that the additional benefits paid for more children are meager, that no solid evidence backs up the claim that welfare encourages illegitimacy, and that reducing benefits would be cruel—especially to the children.

A third problem is geographical disparities in benefits. It is widely believed (though not conclusively proven) that many poor families migrated from the South to northern cities because of the more generous welfare benefits available there. This placed an enormous financial burden on these cities. Finally, the tedious case-by-case approach of AFDC, with its cumbersome bureaucracy and mountains of detailed regulations, seems to frustrate all parties concerned.

Another welfare program that burgeoned in the 1970s and was cut back in the 1980s and 1990s is *Food Stamps*, under which poor families are sold stamps which they can exchange for food. The dollar amount of the stamps they receive, and how much they pay for them, depends on the family's income. The poorer

the family, the less it must pay for the stamps. Headlines were made in 1993 when it was reported that a stunning 10 percent of all Americans were receiving Food Stamps. That percentage has fallen only slightly since, and the program now costs the federal government about $26 billion per year.

In addition, the government provides many of the poor with a number of important goods and services, either at no charge or at prices that are well below market levels. Medical care under the Medicaid program and subsidized public housing are two notable examples.[5] These programs significantly enhance the living standards of the poor. However, most of them offer benefits that decline as family income rises. Taken as a whole, all the antipoverty programs may actually put a poor family in a position where it is *worse* off if its earnings *rise*—an effective marginal tax rate of over 100 percent. When this occurs, there is a powerful incentive not to work.

WELFARE REFORM AND THE TRADE-OFF

The debate over welfare reform is a good illustration of the two questions that are central to the trade-off between equality and efficiency. First come value judgments: How much equality should society buy? Advocates of cutbacks in welfare believe the government is spending too much money to reduce inequality, and so America should move from a point like *c* to a point like *b* in Figure 17-5. Supporters of welfare argue that we are spending too little and allowing excessive amounts of inequality to persist.

The second question pertains to means more than to ends. Critics of welfare argue that the current system is a terribly inefficient way to redistribute income, for reasons like those we have just discussed. It is more like *abcde* in Figure 17-5 than *ABCDE*. Even defenders of welfare agree that it is an inefficient program. But they doubt that the disincentives are as serious as the critics claim.

THE NEGATIVE INCOME TAX

How can we do the job better? Can we design a simple structure that would get income into the hands of the poor without destroying their incentives to work? The solution suggested most frequently by economists is called the *negative income tax (NIT)*.

Table 17-4 illustrates how an NIT works. A particular NIT plan is defined by picking two numbers: a minimum income level below which no family is allowed to fall (the "guarantee"), and a rate at which benefits are "taxed away" as income rises. The table considers a plan with a $6,000 guaranteed income and a 50 percent tax rate. Thus, a family with no earnings (top row) would receive a $6,000 payment (a "negative tax") from the government. A family earning $2,000 (second row) would have the basic benefit reduced by 50 percent of its earnings. Thus, since half its earnings is $1,000, it would receive $5,000 from the government plus the $2,000 earned income for a total income of $7,000.

Notice in Table 17-4 that, with a 50 percent tax rate, the increase in total income as earnings rise is always half of the increase in earnings. Thus, there is always some incentive to work. Notice also that there is a level of income at which benefits cease—$12,000 in this example. This "break-even" level is not a

[5]The *Medicaid* program pays for the health care of low-income people, whereas *Medicare* is available to all elderly people, regardless of income.

TABLE 17-4

ILLUSTRATION OF A NEGATIVE INCOME TAX PLAN

Earnings	Benefits Paid	Total Income
$ 0	$6,000	$ 6,000
2,000	5,000	7,000
4,000	4,000	8,000
6,000	3,000	9,000
8,000	2,000	10,000
10,000	1,000	11,000
12,000	0	12,000

third number that policymakers can select freely. Rather, it is dictated by the choices of the guarantee and the tax rate. In our example, $6,000 is the maximum possible benefit, and benefits are reduced by 50 cents for each $1 of earnings. Hence, benefits will be reduced to zero when 50 percent of earnings is equal to $6,000—which occurs when earnings are $12,000. The general relation is:

$$\text{Guarantee} = \text{Tax rate} \times \text{Break-even level}$$

The fact that the break-even level is completely determined by the guarantee and the tax rate creates a vexing problem. To make a real dent in the poverty problem, the guarantee must come fairly close to the poverty line. But then, any moderate tax rate will push the break-even level way above the poverty line. This means that families who are not considered "poor" (though they are certainly not rich) will also receive benefits. For example, a low tax rate of 33⅓ percent means that some benefits are paid to families whose income is as high as three times the guarantee level.

The solution seems obvious: Raise the tax rate to bring the guarantee and the break-even level closer together. But then the incentive to work shrinks, and with it the principal rationale for the NIT in the first place. So the NIT is no panacea. Difficult choices must still be made.

THE NEGATIVE INCOME TAX AND WORK INCENTIVES

For people now receiving welfare, the NIT would increase work incentives substantially. However, we have just seen that it is virtually inevitable that a number of families who are now too well-off to collect welfare would become eligible for NIT payments. For these people, the NIT imposes work disincentives, both because it provides them with more income and because it subjects them to the relatively high NIT tax rate, which reduces their after-tax wage rate.[6] Government-sponsored experiments in the 1960s found that recipients of NIT benefits did in fact work less than nonrecipients, but only slightly.

Largely because of its superior work incentives, economists believe that an NIT is a more efficient way to redistribute income than the existing welfare system. In terms of Figure 17-5, the NIT is curve *ABCDE*, while the present system

[6]For a review of income and substitution effects in labor supply analysis, refer to Chapter 16, pages 379–381.

is curve *abcde*. If this view is correct, then replacing the current welfare system with an NIT would lead to both more equality *and* more efficiency. But this does not mean that equalization would become costless. The curve *ABCDE* still slopes downward—by increasing equality, we still diminish the GDP.

Actually, however, adopting an NIT would not be as great a change in social policy as some people imagine. The reason is that we already have one—or perhaps two! Specifically, the Food Stamp program functions very much like an NIT because Food Stamps are used like cash in many neighborhoods. Similarly, a feature of the income tax code called the Earned Income Tax Credit (EITC) resembles an NIT for the working poor and near-poor. The EITC was made substantially more generous in 1993. Then, in 1995 and 1996, the president and the Republican majority in Congress engaged in a pitched battle over cutting it back.

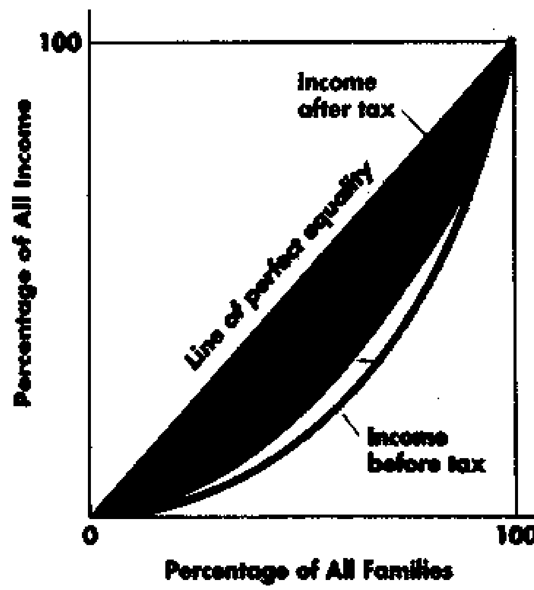
THE PERSONAL INCOME TAX

If we take the broader view that society's objective is not just to eliminate poverty but to reduce income disparities, then the fact that many nonpoor families would receive benefits under an NIT is perhaps not a serious drawback. After all, unless the plan is outlandishly generous, these families will still be well below the average income. Still, the NIT is largely thought of as an antipoverty program, not as a tool for general income equalization.

By contrast, the federal personal income tax *is* thought to be a means of promoting equality. Indeed, it is probably given far more credit for this than it actually deserves. Because the income tax is *progressive*, incomes *after* tax are distributed more equally than incomes *before* tax.[7] This point is illustrated by the two Lorenz curves in Figure 17-6. These curves, however, are not drawn accurately

FIGURE 17-6 THE EFFECT OF PROGRESSIVE INCOME TAXATION ON THE LORENZ CURVE

Since a progressive income tax takes proportionately more income from the rich than from the poor, it reduces income inequality. Graphically, this means that society's Lorenz curve shifts in the manner shown here. The magnitude of the shift, however, is exaggerated to make the graph more readable. In reality, the income tax has only a small effect on the Lorenz curve.



[7]For definitions of progressive, proportional, and regressive taxes, see Chapter 20, page 470.

to scale. If they were, they would lie almost on top of each other because research shows that the degree of equalization attributable to the tax is rather modest.

DEATH DUTIES AND OTHER TAXES

Taxes on inheritances and estates levied by both the state and the federal governments are another equalizing feature of our tax system. And in this case they seem clearly aimed at limiting the incomes of the rich, or at least at limiting their ability to transfer this largesse from one generation to the next. But the amount of money involved is too small to make much difference to the overall distribution of income. Total receipts from estate and gift taxes by all levels of government are well under 1 percent of total tax revenues.

There are many other taxes in the U.S. system, and most experts agree that the remaining taxes as a group—including sales taxes, payroll taxes, and property taxes—are decidedly regressive. On balance, the evidence seems to suggest that:

The U.S. tax system as a whole is only slightly progressive.

POLICIES TO COMBAT DISCRIMINATION

The policies we have just considered are all based on taxes and transfer payments—on moving dollars from one set of hands to another. This has not been the approach used to fight discrimination. Instead, governments have made it *illegal* to discriminate.

Perhaps the major milestone in the war against discrimination was the *Civil Rights Act of 1964*, which outlawed many forms of discrimination and established the *Equal Employment Opportunity Commission (EEOC)*. When you read a want ad in which a company asserts it is "an equal opportunity employer," the firm is proclaiming its compliance with this and related legislation.

Originally, it was thought that the problem could be attacked by outlawing discrimination in rates of pay and in hiring standards—and by devoting resources to enforcement of these provisions. While progress in reducing discrimination by race and sex undoubtedly was made between 1964 and the early 1970s, many people felt the pace was too slow. One reason was that discrimination in the labor market proved to be more subtle than was first thought. Officials rarely could find proof that unequal pay was being given for equal work, because determining when work was "equal" turned out to be a formidable task.

So a new approach was added. Firms and other organizations with suspiciously small representations of minorities or women in their work forces were required not just to end discriminatory practices, but also to demonstrate that they were taking **affirmative action** to remedy this imbalance. That is, they had to *prove* that they were making efforts to locate members of minority groups and females and to hire them if they proved to be qualified.

AFFIRMATIVE ACTION refers to active efforts to locate and hire members of underrepresented groups.

This approach to fighting discrimination was controversial from the start, and remains so to this day. (See the box on the next page.) Critics claim that affirmative action really means quotas and compulsory hiring of unqualified workers simply because they are black or female. If so, it exacts a toll on economic efficiency. Proponents counter that without affirmative action, discriminatory



HAS AFFIRMATIVE ACTION OUTLIVED ITS USEFULNESS?

Affirmative action became a hot political issue in 1995, as conservative politicians reacted to what they perceived to be one of the chief grievances of the "angry white male." A number of Republican presidential hopefuls, including Senators Bob Dole of Kansas and Phil Gramm of Texas and California Governor Pete Wilson, declared that affirmative action had outlived its usefulness. It was time, they said, to rely on "race blind" standards that judge each person on his or her individual merits.

While no federal laws were changed, Wilson abolished several affirmative action programs at the University of California and sought other curbs on affirmative action through a statewide referendum. President Clinton ordered a comprehensive review of federal programs that favored minorities in such matters as hiring and awarding contracts. While a few programs were cut back or eliminated, the review generally concluded that the United States was still so far from being a "color blind" society that affirmative action was still needed. Finally, an important 1995 Supreme Court ruling in *Adarand v. Pena* set new and tougher standards for federal affirmative action programs.

employers would simply claim they were unable to find qualified minority or female employees.

The difficulty revolves around the impossibility of deciding on *purely objective criteria* who is "qualified" and who is not. What one person sees as government coercion to hire an unqualified applicant to fill a quota, another sees as a discriminatory employer being forced to mend his or her ways. Nothing in this book—or anywhere else—will teach you which view is correct in any particular instance.

The controversy over affirmative action is another excellent example of the trade-off between equality and efficiency. There is no doubt that giving more high-paying jobs to members of minority groups and to women would make the distribution of income more equal. Supporters of affirmative action seek this result. But if affirmative action disrupts industry and requires firms to replace "qualified" white males with other, "less qualified" workers, the nation's productivity may fall. Opponents of affirmative action are disturbed by these potential efficiency losses. How far should these programs be pushed? A good question, but one without a good answer.

POSTSCRIPT ON THE DISTRIBUTION OF INCOME

Now that we have completed our analysis of the distribution of income, it may be useful to see how it all relates to our central theme: *What does the market do well, and what does it do poorly?*

We have learned that a market economy uses the marginal productivity principle to assign an income to each individual. In so doing, the market attaches high prices to scarce factors and low prices to abundant ones, and therefore guides firms to make *efficient* use of society's resources. This is one of the market's great strengths.

However, by attaching high prices to some factors and low prices to others, the market mechanism often creates a distribution of income that is quite unequal. Some people wind up fabulously rich while others wind up miserably poor. For this reason, the market has been widely criticized for centuries for doing a rather poor job of distributing income in accord with commonly held notions of *fairness* and *equity*.

On balance, most observers feel that the criticism is justified: The market mechanism is extraordinarily good at promoting efficiency but not very good at promoting equality. As we said at the outset, the market has both virtues and vices.

SUMMARY

1. President Lyndon Johnson declared the War on Poverty in 1964, and within a decade the fraction of families below the official **poverty line** had dropped substantially. However, the poverty population has risen since the late 1970s.
2. The difficulty in agreeing on a sharp dividing line between the poor and the nonpoor leads one to broaden the problem of poverty into the problem of inequality in incomes.
3. In the United States today, the richest 20 percent of families receive about 47 percent of the income, while the poorest 20 percent of families receive just over 4 percent. These numbers have changed little on balance since World War II, although inequality has increased noticeably in the 1980s and 1990s. The U.S. income distribution appears to be more unequal than those of most other industrial nations.
4. Individual incomes differ for many reasons. Differences in native ability, in the desire to work hard and to take risks, in schooling and experience, and in inherited wealth all account for income disparities. Discrimination also plays a role. All of these factors, however, explain only part of the inequality that we observe. A portion of the rest is due simply to good or bad luck, and the balance is unexplained.
5. Prejudice against a minority group may lead to discrimination in rates of pay, or to segregation in the workplace, or to both. However, discrimination may also arise even when there is no prejudice (this is called **statistical discrimination**).
6. There is a *trade-off between the goals of reducing inequality and enhancing economic efficiency*: Policies that help on the equality front normally harm efficiency, and vice versa. This is one of the **Ideas for Beyond the Final Exam.**
7. Because of this trade-off, there is an *optimal degree of inequality* for any society. Society finds this optimum in the same way that a consumer decides how much to buy of different commodities: The trade-off tells us how costly it is to "purchase" more equality, and preferences then determine how much should be "bought." However, since people differ in their value judgments about the importance of equality, there is disagreement over the ideal amount of equality.
8. Whatever goal for equality is selected, society can gain by using more efficient redistributive policies because these policies let us buy any given amount of equality at a lower price in terms of lost output. Economists claim, for example, that a *negative income tax* is preferable to our current welfare system on these grounds.
9. But the negative income tax is no panacea. Its primary virtue lies in the way it preserves incentives to work. But if this is done by keeping the tax rate low, then either the minimum guaranteed level of income will have to be low or many nonpoor families will become eligible to receive benefits.
10. The goal of income equality is also pursued through the tax system, especially through the progressive federal income tax and death duties. But other taxes are typically regressive, so the tax system as a whole is only slightly progressive.
11. **Economic discrimination** has been attacked by making it illegal, not through the tax and transfer system. But simply declaring discrimination to be illegal is much easier than actually ending discrimination. The trade-off between equality and efficiency applies once again: Strict enforcement of **affirmative action** will certainly reduce discrimination and increase income equality, but it may do so at a cost in terms of economic efficiency.

KEY TERMS

Poverty line
Absolute and relative concepts of poverty
Lorenz curve
Economic discrimination
Statistical discrimination
Optimal amount of inequality
Trade-off between equality and efficiency
Aid to Families with Dependent Children (AFDC)
Food Stamps
Negative income tax (NIT)
Civil Rights Act of 1964
Equal Employment Opportunity Commission (EEOC)
Affirmative action

QUESTIONS FOR REVIEW

1. Discuss the "leaky bucket" analogy (page 419) with your classmates. What maximum amount of income would you personally allow to leak from the bucket in transferring money from the rich to the poor? Explain why people differ in their answers to this question.
2. Continuing the leaky bucket example, explain why economists believe that replacing the present welfare system with a negative income tax would help reduce the leak.
3. Suppose you were to design a negative income tax system for the United States. Pick a guaranteed income level and a tax rate that seem reasonable to you. What break-even level of income is implied by these choices? Construct a version of Table 17-4 (page 422) for the plan you have just devised.
4. Following is a complete list of the distribution of income in Disneyland. From these data, construct a Lorenz curve for Disneyland.

Donald Duck	$100,000
Mickey Mouse	172,000
Minnie Mouse	68,000
Pluto	44,000
Ticket taker	16,000

   How different is this from the Lorenz curve for the United States (Figure 17-2 on page 409)?
5. Suppose the War on Poverty were starting anew and you were part of a presidential commission assigned the task of defining the poor. Would you choose an absolute or a relative concept of poverty? Why? What would be your specific definition of poverty?
6. Discuss the concept of the "optimal amount of inequality." What are some of the practical problems in determining how much inequality really is optimal?
7. Why do you think the distribution of income has grown more unequal during the 1980s and 1990s?
8. A number of conservative politicians and economists advocate replacing the progressive income tax with a "flat tax" that would apply the same, low tax rate to all income above a certain exempt amount. One argument against making this change is that the distribution of income has grown much more unequal since the 1970s. Does the evidence support that view? Is it a decisive argument against a flat tax? How is the trade-off between equality and efficiency involved here?



PART V

The Government and the Economy

CHAPTER 18

LIMITING MARKET POWER: REGULATION OF INDUSTRY

. . . We believe that competition, for all its imperfections, is proving itself superior to regulation as a means of serving the public interest.
Bailey, Graham, and Kaplan, *Deregulating the Airlines*, 1985

Because the market may not function ideally in monopolistic or oligopolistic industries, governments often intervene. In the United States, such intervention normally follows two basic patterns. Antitrust laws, the subject of the next chapter, seek to prohibit acquisition of monopoly power and to ban certain monopolistic practices. In addition, some firms are subjected to **regulation,** which constrains their pricing policies and other decisions.

REGULATION of industry is a process established by law that restricts or controls some specified decisions made by the affected firms. Regulation is usually carried out by a special government agency assigned the task of administering and interpreting the law. That agency also acts as a court in enforcing the regulatory laws.

Yet, despite the good intentions of its designers, the regulatory mechanism, particularly in the form it took before the 1980s, was criticized for costing the consuming public dearly rather than protecting its interests. This chapter will explain the problems and the steps, many of them suggested by simple economic theory, taken to remedy them.

In evaluating these changes, we should keep in mind that the purpose of regulation is, in essence, to prevent the market mechanism from working as it would if it were left unhampered. We saw in Chapter 13 that various market activities may well need improvement. But many observers conclude that government regulators with the power to intervene do not generally stop where the logic of Chapter 13 indicates they should. Rather, they are all too likely to interfere with the market mechanism in places where it is best left alone.

PUZZLE: INDUSTRY OPPOSITION TO DEREGULATION

An observer who knew nothing about regulated industries might expect that firms would welcome deregulation. After all, regulations curb companies' freedom of decision making in many ways.

Yet many industries—and their unions—have bitterly fought deregulation. Take the trucking industry, for example. When Congress proposed doing away with state rules that regulated truck rates, routes, and services, there were loud protests from the Teamsters Union, which strongly opposed the move. Later, we

will discuss some reasons for this opposition. But already we may surmise that regulation sometimes may, inadvertently or deliberately, serve the interests of regulated industries rather than making life harder for them.

MONOPOLY, REGULATION, AND NATIONALIZATION

As we saw in Chapter 11, a number of industries traditionally operate as monopolies. These include postal services and transportation and utility companies. Since there is little or no competition to protect consumers from monopolistic exploitation in these cases, substitute forms of protection from excessive prices and restricted outputs have been adopted.

NATIONALIZATION is the acquisition of a private firm by the government.

PRIVATIZATION is the return of a government firm to private ownership.

Until recently, most of western Europe had chosen **nationalization** as its solution, in which the state owns and operates certain monopolistic industries. However, much of the world has had second thoughts about this approach in recent years. The last decade has witnessed an outburst of **privatization**—sales to private owners of firms formerly owned by government—from eastern Europe to Latin America and in much of the British Commonwealth. This process is still going on.

The United States is traditionally more reluctant to have government-run businesses. Yet even here it has happened to some degree. Most cities now operate their own public transport systems; public corporations run the post office and much of the nation's passenger railroad system; the Tennessee Valley Authority is a major experiment in electricity supplied by a public agency.

Since nationalization is the exception rather than the norm in the United States, the government depends on regulatory agencies to control privately owned monopolies. Both federal and state governments have created a large number of agencies that regulate prices, standards of service, provisions for safety, and a variety of other aspects of the operations of telephone companies, radio and television stations, electric utilities, airlines, trucking companies, and firms in many other industries—all of which remain in private ownership. Many of these industries are not pure monopolies, but include firms that nevertheless are believed to possess so much market power that their regulation is considered to be in the public interest. In some other countries, nationalized firms that have undergone privatization have immediately been subjected to supervision by a regulatory agency, much like their U.S. counterparts.

WHAT IS REGULATED? BY WHOM?

The regulatory agencies in the United States can be divided, roughly, into two classes: those that limit the market power of regulated firms and those devoted to consumer and worker protection and safety. In a recent count, at least 14 federal regulatory agencies dealt with restraint of market power and about 30 handled issues such as environmental protection and product safety. A primary example of the latter is the Food and Drug Administration (FDA), whose tasks are twofold: protecting the public from harmful, impure, infected, or adulterated foods, drugs, and cosmetics, and preventing the mislabeling of these products. Similarly, the U.S. Department of Agriculture supervises the packing and grading of meats and poultry going into interstate commerce.

The federal government also regulates the safety of automobiles and mines and the use of such substances as dangerous pesticides. Such regulation affects an enormous proportion of the nation's economic activity, including the drug indus-

try, agriculture, auto manufacturing, and the chemical and power industries. Virtually every manufacturing industry is affected by environmental regulations.

Regulations designed to limit market power affect industries that together provide perhaps 10 percent of the GDP of the United States. The list includes telecommunications, railroads, electric utilities, and oil pipelines.

A BRIEF HISTORY OF REGULATION

Regulation of industry in the United States first began when indignation over abuse of market power by the nation's railroads led to the establishment of the Interstate Commerce Commission (ICC) in 1887. In particular, there was a public outcry over the support the railroads gave to John D. Rockefeller, Sr., in the battle of his Standard Oil Company against its rivals. This, along with other abuses by the railroads, invited government intervention. But for several decades afterward, there was little attempt to expand regulation to other industries. Then the Federal Power Commission (FPC) was established in 1920 and the Federal Communications Commission (FCC) in 1934; a substantial proportion of the remaining regulatory agencies were also formed during the 1930s as part of President Franklin D. Roosevelt's New Deal.

Today, several regulatory agencies of the federal government control prices. Until it was abolished in 1995, the ICC regulated railroads, barges, pipelines, and some categories of trucking. Its regulatory tasks are now performed by the Department of Commerce. The FCC regulates broadcasting and telecommunications. The Federal Energy Regulatory Commission (FERC) regulates interstate transmission of electric power and sales of natural gas. The Securities and Exchange Commission (SEC) regulates sales of securities (stocks). A number of agencies, led by the Federal Reserve System, control banking operations. The work of these agencies is complemented by a variety of state agencies that regulate intrastate activities.

Economists have long questioned the effectiveness and desirability of regulation. But such questions were not raised seriously outside of academia until the mid-1970s, when Congress enacted laws that limited the powers of regulatory agencies. Several industries, such as airlines and trucking, were substantially "deregulated"—that is, most of the powers of the regulatory agencies were eliminated. In other industries, such as railroads and telecommunications, the rules have been changed to give regulated firms considerably more freedom in their decision making. This process is still under way. Indeed, Congress engaged in a major—and highly charged—debate over telecommunications regulation in 1995, ending in a major piece of legislation in 1996.

From the 1970s through the 1990s, Presidents Ford, Carter, Reagan, Bush, and Clinton all concluded that the economy was overregulated and that this imposed unnecessary costs on consumers. Deregulation began in earnest in the last few years of the Carter administration. In 1978, Congress passed an act ending regulation of passenger air transportation. In the period since then, regulatory control over truck and rail transportation has been curtailed sharply.

In telecommunications, AT&T's monopoly was ended and the firm itself broken up under the terms of settlement of an antitrust case. The local telephone companies that were separated from AT&T continue to be regulated, though the regulatory rules are now being rewritten in Congress and in various state and local arenas. As this is being written, Congress is seeking additional ways to reduce regulation of business.

WHY REGULATION?

Economists recognize a number of reasons that sometimes justify the regulation of an industry.

NATURAL MONOPOLY AND SCALE ECONOMIES

ECONOMIES OF SCALE are savings that are acquired through increases in quantities produced.

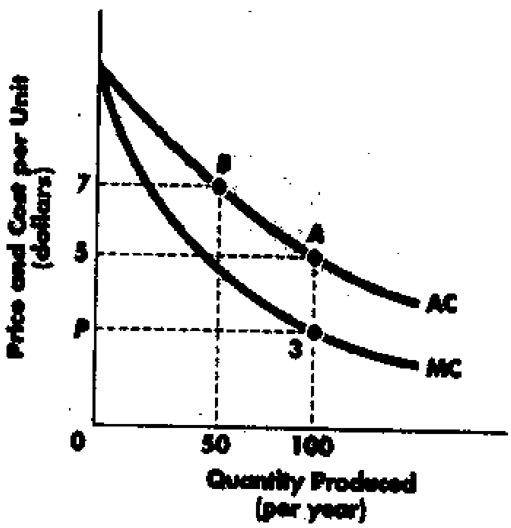
As we learned in Chapter 11, one main reason for regulation of industry is the phenomenon of *natural monopoly*. In some industries, it is apparently far cheaper to have production carried out by one firm rather than by many different firms. One reason for this is the presence of economies of large-scale production. An example of such **economies of scale** is a railroad track. The total track cost of carrying 100 trains a day is hardly higher than carrying one. Here is a case in which savings are made possible by expanding the volume of an activity—a case of economies of scale. As we saw in Chapter 7, scale economies lead to an average cost curve that goes downhill as output increases (see Figure 18-1). This means that a firm with a large output can cover its costs at a price lower than a firm whose output is smaller. In Figure 18-1, point *A* represents the larger firm whose AC is $5 while *B* is the smaller firm with an AC of $7.

ECONOMIES OF SCOPE are savings that are acquired through simultaneous production of many different products.

A single, large firm may also have a cost advantage over a group of small firms because it is sometimes cheaper to produce *a number of different commodities together* rather than making each separately in a different firm. Savings made possible by simultaneous production of many different products by one firm are called **economies of scope**. An example of economies of scope is the manufacture of both cars and trucks by the same producer. The techniques employed in producing both commodities are similar, and this can provide a cost advantage to firms that produce both.

In industries where there are great economies of scale *and* scope, society will obviously incur a significant cost penalty if it insists on maintaining a large number of small, and therefore costly, firms. Moreover, in the presence of strong economies of scale and economies of scope, society *will not be able to preserve free*

FIGURE 18-1 MARGINAL-COST PRICING UNDER ECONOMIES OF SCALE



Economies of scale imply that the average cost (AC) curve is declining and, therefore, that the marginal cost (MC) curve is below the average cost curve. If, for example, the regulator forces the firm to produce 100 units and charge a price equal to its marginal cost ($3 per unit), then the firm will take in $300 in revenues. But, since its average cost at 100 units is $5 per unit, its total cost will be $500, and the firm will lose money.

competition, even if it wants to. The large, multiproduct firm will have so great a cost advantage over its rivals that the small firms simply will be unable to survive. We say in such a case that free competition is *not sustainable.*

Where monopoly production is cheapest, and where free competition is not sustainable, the industry is a natural monopoly. When monopoly is cheaper, society may not want to have competition; if free competition is not sustainable, it will not even have a choice in the matter.

But even if society reconciles itself to monopoly, it will generally not want to let the monopoly firm do whatever it wants to with its market power. Therefore, it will consider either regulation of the company's decisions on matters such as prices or nationalization of monopoly firms.

UNIVERSAL SERVICE AND RATE AVERAGING

A second reason for regulation is the desire for "universal service," that is, the availability of service at "reasonable prices," even to small communities where the limited scale of operations may make costs extremely high. In such cases, regulators may encourage or require a public utility to serve some consumers at a financial loss. But a loss on some sales is financially feasible only when the firm is permitted to make up for it by obtaining higher profits on its other sales.

CROSS-SUBSIDIZATION means selling one product at a loss, which is balanced by higher profits on another product.

This so-called *rate averaging* of gains and losses, also referred to as **cross-subsidization,** is possible only if the firm is protected from price competition and free entry of new competitors in its more profitable markets. If no such protection is provided, potential competitors will sniff out the profit opportunities in the markets where service is supplied at prices well above cost. Many new firms will enter the business and drive prices down in those markets—a practice referred to as *cream skimming.* The entrants choose to enter only the profitable markets and skim away the cream of the profits for themselves, leaving the unprofitable markets (the skimmed milk) to the supplier who had attempted to provide universal service. This phenomenon is one reason why regulatory rules, until recently, made it very difficult or impossible for new firms to enter when and where they saw fit.

Airlines and telecommunications are two industries in which these issues have arisen. In both cases, it was feared that without regulation of entry and rates, or special subsidies, less populous communities would effectively be isolated, losing their airline services and obtaining telephone service only at cripplingly high rates. Many economists question the validity of this argument for regulation, which, they say, calls for hidden subsidies of rural consumers by everyone else. The airline deregulation act provided for government subsidies to help small communities attract airline service. In fact, this market has since been taken over to a considerable extent by specialized "commuter" airlines flying much smaller aircraft than the major airlines, which have withdrawn from many such routes.

A similar issue affects the U.S. Postal Service, which charges the same price to deliver a letter anywhere within the United States, regardless of the distance or the special difficulties and costs of a particular route. To maintain this pricing scheme, the law must protect the Postal Service from direct competition in many of its activities; otherwise, its extreme form of uniform pricing would soon deprive it of its most profitable routes.

We conclude that the goal of universal service leads to regulatory control of entry and exit and not just to control of prices.

DESTRUCTIVE COMPETITION

A third reason for regulation is to help prevent *self-destructive competition*, made possible, for example, by economies of scale. In an industry such as railroading, equipment—including roadbeds, tracks, switching facilities, locomotives, and cars—is extremely expensive. Suppose that two railroads are competing for some limited business that happens to be insufficient to use their total facilities to anything near capacity. That is, to meet this low level of consumer demand, each railroad runs, say, only 40 percent of the trains that it could schedule on its tracks.

The management of each railroad will feel that, with its unused capacity, any business will be worthwhile, provided that it covers more than its short-run marginal costs—fuel, labor, and expenses other than plant and equipment. If the short-run marginal cost of shipping an additional ton of, say, coal is $5, then either railroad will be happy to lure coal-shipping customers away from the other at a price of, say, $7 per ton. Even though such prices may not cover the entire cost of track and equipment, each ton of business will put the railroad $2 ahead of where it would have been without it.

Thus, even if the new business only pays for its own marginal cost and a little more, it seems financially desirable. But the temptation to accept business on such terms will drive both firms' prices down toward their marginal costs, and, in the process, both railroads are likely to go broke. If no customer pays for the track, the roadbed, and the equipment, the railroad simply will be unable to stay in business.

For this reason, it is argued, regulation of rates can be sensible, even in industries subject to competitive pressures, simply to protect the industries from themselves. Without this regulation, self-destructive competition could end up sinking those industries financially, and the public would thereby be deprived of vital services. There are those who believe that this problem now besets the major U.S. airlines, most of which have been sustaining losses, some of them huge, for a considerable number of years. Some now have returned to profitability, but earnings are generally below what suppliers of capital can obtain in competitive industries. That is, earnings still do not cover the opportunity cost of the firms' investments, so they continue to generate negative economic profits. (For a review of the concept, see Chapter 9, pages 225–226.)

PROTECTION AGAINST MISINFORMATION

A final reason for regulation is the danger that consumers will be misinformed or cheated; that consumers, company employees, or the environment will be threatened by unscrupulous sellers; or that even conscientious sellers will be forced to match the questionable practices of less scrupulous rivals. Protection of this nature is the province of the second type of regulatory agency described earlier.

WHY REGULATORS SOMETIMES RAISE PRICES

Regulation sometimes forces consumers to pay more than they would pay in its absence. One of the most widely publicized examples illustrating the tendency of regulation to push rates upward was the difference in airfares between San Francisco and Los Angeles and those between Washington, D.C., and New York City before deregulation. The federal government never regulated the California

fare (since the flight is entirely within the state), whereas it did control the interstate flight between New York and Washington, D.C. The distance of the California trip is nearly twice as great as the East Coast trip, and neither route is sparsely traveled nor beset by any other noteworthy features that would make for substantial differences in cost per passenger mile. Yet at the time of deregulation, fares were a little over $40 for the long California journey and a little over $50 for the short Washington to New York trip.

Regulators sometimes push for higher prices when they want to prevent the demise of any existing firms in an industry. We saw earlier that strong economies of scale and scope may make it impossible for a number of firms to survive. The largest firm in an industry may have such a big cost advantage over its competitors that it will be able to drive them out of the market while still operating at prices that are profitable. Most observers applaud low prices and price cuts that reflect such cost advantages. However, a firm that wants the market for itself may conceivably engage in price cutting even when such cuts are not justifiable in terms of cost.

Setting a price below the pertinent cost may not reduce the overall profits of a regulated firm because regulation often imposes an upper limit on the amount of profit that a firm is permitted to earn. To see the connection, consider a regulated firm that produces two commodities, A and B, and that is forced to set each price below its profit-maximizing level in order to limit profits to the allowable ceiling. The firm may be able, without loss of profit, to cut the price of Commodity A even below its marginal cost and persuade the regulators to let it make up for any resulting decrease in profit by a sufficient rise in the price of Commodity B. In this case, we say that the firm has instituted a *cross-subsidy* from the consumers of Commodity B to the consumers of Commodity A. That is, consumers of Commodity B pay excessive amounts for their purchases to make up for the deficit that the firm suffers in sales of Commodity A.

Why would any firm want to do this? Suppose that Commodity A is threatened by competition, while Commodity B has no competitors on the horizon. Then a cross-subsidy from B to A may permit a cut in the price of A sufficient to prevent the entry of the potential competitors of A or even to drive some current competitors out of the field. The fear by the Department of Justice of such a cross-subsidy of telephone equipment by the monopoly local telephone companies, which formerly were subsidiaries of AT&T, was one of the elements underlying the decision to break up the single-firm telephone system. After all, one way to prevent a cross-subsidy from Product B to Product A is to require B and A to be produced by two different firms.

But regulation sometimes goes beyond the prevention of cross-subsidies. Firms that feel they are hurt by competitive pressures will complain to regulatory commissions that the prices charged by their rivals are "unfairly low." The commission, afraid that unrestrained pricing will reduce the number of firms in the industry, then attempts to "equalize" matters by imposing price floors that permit all the firms in the industry to operate profitably, even those that operate inefficiently and incur costs far higher than their competitors' costs.

Many economists maintain that this approach to pricing is a perversion of the idea of competition. The virtue of competition is that, where it occurs, firms force one another to supply consumers with products of high quality at *low* prices. Any firm that cannot do this is driven out of business by market forces. An arrangement may allow firms to coexist only by *preventing* them from competing with one another; this preserves the appearance of competition but destroys its substance.

A PROBLEM OF MARGINAL-COST PRICES

Government agencies are sometimes assigned the task of setting or at least influencing the prices of regulated firms. The acrimonious debate over the proper levels for those prices has filled hundreds of thousands of pages of regulatory-hearing records and has involved literally hundreds of millions of dollars of expenditures in fees for lawyers, expert witnesses, and research. The question has been what constitutes the proper formula to set these prices.

Where it is feasible, most economists favor setting price equal to marginal cost because, as was shown in Chapter 10, this pricing policy provides the incentive for firms to produce output quantities that serve consumers' wants most efficiently. However, a serious practical problem prevents the use of the principle of marginal-cost pricing in many industries. The problem is that, in many regulated industries, the firms would go bankrupt if all prices were set equal to marginal costs.

This seems a startling conclusion, but it follows inescapably from three simple facts:

FACT 1. In many regulated industries, there are significant economies of large-scale production. As we pointed out earlier, economies of scale are one of the main reasons why certain industries were regulated in the first place.

FACT 2. In an industry with economies of scale, the long-run average cost curve is downward sloping. This means that long-run average cost falls as the quantity produced rises, as illustrated by the AC curve in Figure 18-1 on page 434. Fact 2 is something we learned back in Chapter 7 (pages 159–161 and 167–169). The reason, to review briefly, is that total costs must double if all input quantities are doubled. But, where there are economies of scale, output will *more* than double if all input quantities are doubled. Since average cost (AC) is simply total cost (TC) divided by quantity (Q), AC = TC/Q must decline when all input quantities are doubled.

FACT 3. If average cost is declining, then marginal cost must be below average cost. This fact follows directly from one of the general rules relating marginal and average data that were explained in the appendix to Chapter 8. Once again, the logic is simple enough to review briefly. If, for example, your average quiz score is 90 percent but the next quiz pulls your average down to 87 percent, then the grade on this most recent test (the marginal grade) must be below both the old and the new average quiz scores. That is, it takes a marginal grade (or cost) that is below the average to pull the average down.

Putting these three facts together, we conclude that in many regulated industries, marginal cost (MC) will be below average cost, as depicted in Figure 18-1. Now suppose that regulators set the price at the level of marginal cost. Since P equals MC, P must be below AC and the firm must lose money, so P equals MC is simply not an acceptable option. What, then, should be done? One possibility is to nationalize the industry, set price equal to marginal cost, and make up for the deficit out of public funds. Nationalization, however, is not very popular in the United States or, nowadays, in most other countries. (More is said about nationalization at the end of the chapter.)

A second option, which is quite popular among regulators, is to (try to) set price equal to *average cost*. But this method of pricing is neither desirable nor possible to carry out except on the basis of arbitrary decisions. The problem is

that almost no firm produces only a single commodity. Almost every company produces a number of different varieties and qualities of some product, and many produce thousands of different products, each with its own price. Even General Motors, a fairly specialized firm that produces many makes and sizes of cars and trucks, also sells home mortgages and quite a few other things. In a multiproduct firm, we cannot even define AC = TC/*Q*, since to calculate *Q* (total output), we would have to add up all the apples and oranges (and all of the other different items) that the firm produces. But we know that one cannot add up apples and oranges. Since we cannot calculate AC for a multiproduct firm, it is hardly possible for the regulator to require *P* to equal AC for each of the firm's products, though regulators sometimes think they can do so.

THE RAMSEY PRICING RULE

In recent years, economists have been attracted to an imaginative, alternative approach to the problem of pricing in regulated industries that produce many products. This approach derives its name from its discoverer, Frank Ramsey, a brilliant English mathematician who died in 1930 at the age of 26 after making several enduring contributions to both mathematics and economics.

The basic idea of Ramsey's pricing principle can be explained in a fairly straightforward manner. We know that prices must be set *above* marginal costs if a firm with scale economies is to break even, but how much above? In effect, Ramsey argued as follows: The reason that we do not like prices to be above marginal costs is that such high prices distort the choices made by consumers, leading them to buy "too little" of the goods whose prices are set way above MC. Yet, it is necessary to set prices somewhat above marginal costs to allow the firm to survive. Therefore, it makes sense to raise prices *most* above marginal cost where consumers will respond the *least* to such a price increase, that is, where the *elasticity of demand* is lowest so that price rises will create the least distortion of demand. This line of argument led Ramsey to formulate the following rule:

The **RAMSEY PRICING RULE** is a rule for determining prices that promote consumer welfare while covering the producer's cost.

The **Ramsey Pricing Rule:** In a multiproduct, regulated firm in which prices must exceed marginal cost in order to permit that firm to break even, other things being equal, the ratios of *P* to MC should be largest for those of the firm's products whose elasticities of demand are the smallest.

Economists accept this pricing rule as the correct conclusion on theoretical grounds. It has even been proposed for postal and telephone pricing, and the Interstate Commerce Commission (ICC) explicitly decided to adopt the Ramsey principle as its general guide for the regulation of railroad rates.

MODIFIED RAIL REGULATION POLICY AND STAND-ALONE COST CEILINGS

In the regulation of railroads, the ICC has adopted a new approach to regulation explicitly derived from the theory of contestability that we mentioned in Chapter 12 (see pages 295–296). In its decision, the ICC recognized the value of the Ramsey pricing rule as a general guideline for policy. (Excerpts from this ICC decision are quoted in the box on the next page.) But the commissioners felt that it was not practical to calculate statistically and update constantly all the demand elasticity numbers and marginal cost figures that the use of the Ramsey rule requires.

ECONOMIC THEORY IN AN ICC COAL RATE DECISION



In this decision the Interstate Commerce Commission based its determination directly on economic analysis. Stand-alone cost (SAC) is the ceiling imposed on a railroad's prices because no higher prices could be charged in an unregulated competitive market.

> ... [The] stand-alone cost (SAC) test ... is used to compute the rate a competitor in the market-place would need to charge in serving a captive shipper or a group of shippers who benefit from sharing joint and common costs. A rate level calculated by the SAC methodology represents the theoretical maximum rate that a railroad could levy on shippers without substantial diversion of traffic to a hypothetical competing service. It is, in other words, a simulated competitive price....
>
> The theory behind SAC is best explained by the concept of contestable markets. This recently developed economic theory augments the classical economic model of pure competition with a model which focuses on the entry and exit from an industry as a measure of economic efficiency.... The underlying premise is that a monopolist or oligopolist will behave efficiently and competitively where there is a threat of losing some or all of its markets to a new entrant. In other words, contestable markets have competitive characteristics which preclude monopoly pricing.

The applicability of the principles in this decision to rates on other types of freight was reconfirmed in 1993 by the U.S. Court of Appeals (Case No. 88-1114, decided February 9, 1993). The decision continues in force after replacement of the ICC by the Department of Commerce.

SOURCE: Excerpted from Interstate Commerce Commission, "Coal Rate Guidelines, Nationwide," Ex Parte No. 347 (Sub-No. 1), August 3, 1985, p. 10.

Instead the ICC decided to adopt a four-part rule. Its intent is to compel railroads to set the prices they would have set if all of their activities were contestable, that is, as if entry into freight transportation were easy enough to subject the railroads to a perpetual and constant threat of new competition. The four parts of the new rule are:

1. For those types of freight and routes where competition happens to be substantial and effective, the railroads should be deregulated; that is, let market forces do the job of policing the railroads' behavior.
2. Where competition is inadequate, a floor and a ceiling should be set for each and every railroad price, and regulators should leave the railroads free to select any level of price they wish within those bounds.
3. The price floor should be the lowest level to which the price could fall in the long run under perfectly competitive conditions. This provision, in effect, prohibits the railroad from adopting any price below marginal cost. It is designed to provide adequate and defensible protection to a railroad's rivals against any attempt by the railroad at unfair competitive price cutting.
4. The price ceiling should be the cost that a *hypothetical* (that is, imaginary) efficient entrant would have to bear to supply each specific service. In other words, in activities where entry is difficult or impossible, the idea is to prohibit the railroads from charging more than they could get away with if entry were instead easy and cheap. The hypothetical cost figure for the efficient entrant is called the *stand-alone cost* of the service. It is the cost that would be required if an efficient entrant were to supply just the service or group of services in question. This provision is intended to protect the interests of railroad customers, guaranteeing them prices no higher than those that might be charged if the markets were effectively contestable.

Most economists who have studied the issue seem to approve of this new approach to rate regulation, which today guides railroad rates under the Department of Commerce, although there are still some disputes about details.

REGULATION OF PROFIT AND INCENTIVES FOR EFFICIENCY

Many opponents of regulation maintain that it seriously impairs the efficiency of American industry. Government regulation, these critics argue, interferes with the operation of Adam Smith's invisible hand. One source of inefficiency—the seemingly endless paperwork and complex legal proceedings that impede the firm's ability to respond quickly to changing market conditions—is obvious enough. (What to do about this administrative problem is far from obvious, however.)

In addition, economists believe that regulatory interference in pricing decisions adds to economic inefficiency. By forcing prices to be either lower or higher than those that would prevail on a free, competitive market, regulations give consumers the wrong signals and induce them to demand a quantity of the regulated product that is inconsistent with maximization of consumer benefits from the quantity of resources available to the economy. (This resource misallocation issue was discussed in Chapter 13, pages 302–304.)

But there is a third source of inefficiency that may be even more important. It stems from the problem that regulators have of trying to prevent the regulated firm from earning excessive profits, while at the same time (a) offering it financial incentives for maximum efficiency of operation and (b) allowing it enough profit to attract the capital it needs when growing markets justify expansion. From this point of view, it would be ideal if the regulator would just permit the firm to take in that amount of revenue that covers its costs, including the cost of its capital. That is, the firm should earn exactly enough to pay for its ordinary costs plus the normal profit that potential investors could get elsewhere for the same money (the opportunity cost of the money). Thus, if the prevailing rate of return is 10 percent, the regulated firm should recover its expenditures plus 10 percent on its investment and not a penny more or less.

The trouble with such an arrangement is that it removes all incentive for efficiency, responsiveness to consumer demand, and innovation. Such an arrangement eliminates the profit motive for efficiency and good service to consumers. It, in effect, *guarantees* just *one standard rate* of profit to the firm, no more and no less. This is so whether its management is totally incompetent or extremely talented and hard working.

Competitive markets do *not* work this way. While under perfect competition the *average* firm will earn just the opportunity cost of capital, a firm with an especially ingenious and efficient management will do better, and a firm with an incompetent management is likely to go broke. It is the possibility of great rewards and harsh punishments that gives the market mechanism its power to cause firms to strive for high efficiency and productivity growth.

We have strong evidence that where firms are guaranteed fixed returns no matter how well or how poorly they perform, gross inefficiencies are likely to result. For example, many contracts for purchases of military equipment have offered prices calculated on a *cost-plus* basis, meaning that the supplier was guaranteed that its costs would be covered and that, in addition, it would receive some prespecified amount as a contribution to profit. Studies of the resulting performance of cost-plus arrangements have confirmed enormous supplier inefficiencies.

A regulatory arrangement that in effect guarantees a regulated firm its cost plus a "fair rate of return" on its investment obviously has a good deal in common with a cost-plus contract of an unregulated firm. Fortunately, there are also substantial differences between the two, and so regulatory profit ceilings need not always have serious effects on the firm's incentives for efficiency.

For one thing, when a regulated industry is in financial trouble, as is true of the railroads, there is nothing the regulator can do to guarantee a "fair rate of return." If the current return on capital is 10 percent, but market demand for railroading is only sufficient to give it 3 percent at most, the regulatory agency cannot help matters by any act of magic. Even if it grants higher prices to the railroad (or forces the railroad to raise its prices) the result will be to drive even more business away and lower profits further. Thus, the regulated firm is not promised any minimum profit rate, unlike the case of an unregulated firm with a cost-plus contract.

There is a second reason why profit regulation does not work in the same way as does a cost-plus arrangement. Curiously, this is a result of the much-criticized delays that characterize many regulatory procedures. In a number of regulated industries, a proposed change in rates is likely to take a minimum of several months before it gets through the regulatory machinery. Where it is bitterly contested, the resulting hearings before the regulatory commission, the appeals to the courts, and so on are likely to last for years. Rate cases lasting 10 years are not uncommon. This phenomenon, known as *regulatory lag*, is perhaps the main reason that profit regulation has not eliminated all rewards for efficiency and all penalties for inefficiency.

Suppose, for example, that the regulatory commission approves a set of prices calculated to yield exactly the "fair rate of return" to the company, say, 10 percent. If management then invests successfully in new processes that reduce its costs sharply, the rate of return under the old prices may rise to, say, 12 percent. If it takes 3 years for the regulators to review the prices they previously approved and adjust them to the new cost levels, the company will earn a 2-percent bonus as a reward for its efficiency during the 3 years of regulatory lag. Similarly, if management makes a series of bad decisions that reduce the company's return to 7 percent, the firm may well apply to the regulator for some adjustments in prices to permit it to recoup its losses. If the regulator takes 18 months to act, the firm suffers a penalty for its inefficiency. It may be added that where mismanagement is *clearly* the cause of losses, regulators will be reluctant to permit the regulated firm to make up for such losses by rate adjustments. But, in most cases, it is difficult to pinpoint responsibility for a firm's losses.

All in all, those who have studied regulated industries have come away deeply concerned about the effects of regulation upon economic efficiency. Although some regulated firms seem to operate very efficiently, others seem to behave in quite the opposite way.

While regulatory lag does permit some penalty for inefficiency and some reward for superior performance by the regulated firm, the arrangement works only in a rough-and-ready manner. It still leaves the provision of incentives for efficiency as one of the fundamental problems of regulation. How can one prevent regulated firms from earning excessive profits, but also permit them to earn enough to attract the capital they need while still allowing rewards for superior performance and penalties for poor performance?

PRICE CAPS AS INCENTIVES FOR EFFICIENCY

A regulatory innovation designed to prevent monopoly profits while offering incentives for the firm to improve its efficiency is now in use in Great Britain for electricity, telephones, and airport services and in the United States and else-

where for telephone rates; it is under discussion for other regulated industries. The basic idea is for regulation consciously to take advantage of the incentive for efficiency provided by regulatory lag. Under this program, the regulators assign ceilings (*price caps*) for the product prices of the firms they oversee.

However, the price caps—measured in inflation-adjusted *real* terms—are reduced each year at a rate based on the rate of cost reduction (productivity growth) previously achieved by the regulated firm. Thus, if in the future the regulated firm can manage to achieve cost savings (by innovation or other means) greater than those it obtained in the past, the firm's real costs will fall faster than its real prices, and it will be permitted to keep the resulting profits as its reward for its effective cost-reduction program. Of course, for the regulated firm there is a catch. If the firm proves able to reduce its costs by, say, 2 percent per year in real terms, but on the basis of its past record its regulatory price cap is cut 3 percent per year, the firm will lose profits, though consumers will continue to benefit from the cuts in real prices.

Thus, in order to earn economic profits, management is constantly forced to look for ever more economical ways of doing things. This approach clearly gives up any attempt to limit the profit of the regulated firm. But it protects the consumer nonetheless by controlling the firm's prices. Indeed, it makes those prices lower and lower, in real terms.

PRICING OF ACCESS TO "BOTTLENECK" SERVICES

Bottleneck facilities are a widespread problem in regulated industries and have recently grown in importance. Suppose, for example, that several competing railroads need to cross a single bridge over a river to serve the traffic between two geographic points for which they compete. Suppose also that the bridge is owned by one of those railroads. That bridge is called a *bottleneck facility* because none of the competitors can supply its product without using the resource. Both the owner railroad and its rivals must funnel all service through this "bottleneck."

The same situation arises in telecommunications. Most long-distance calls to or from points in Ohio or Illinois, whether carried by AT&T, MCI, or Sprint, must at some point get to a home or business using the local telephone network—"the local loop"—which is owned by Ameritech. Ameritech hopes to get into some of the long-distance business itself. But then it, along with AT&T, MCI, and Sprint, will have to make use of its own bottleneck local loop. In the electric power industry, rival generators of electricity must all use the bottleneck transmission facilities owned by the utility that also generates electricity of its own.

If competition is not to be destroyed in these cases, rivals who do not own the bottleneck facility must clearly be given permission to use the facility after, of course, paying an appropriate price. But what price is appropriate? This has become the key issue in litigation and negotiation with regulatory agencies throughout the industrial world. Clearly, if the price the owner is allowed to charge rivals for use of the bottleneck facility is too high, it can drive the competitors out of business. On the other hand, if regulators force the bottleneck owner to charge too low a price, it will be subsidizing its rivals and will find it difficult or impossible to compete with them.

What price, then, is neither too high nor too low? The answer provided by a group of economists is suggested by analogy with a simpler situation. Suppose, in our railroad example, that two railroads, X and Y, need to use the bottleneck bridge, but the owner is not itself a railroad. In order to provide a "level playing

field" for the two rival railroads, it is necessary for the bridge owner to charge the same bridge toll to Railroad X that it charges to Railroad Y. Then, with both railroads paying the same bridge-use fee, the railroad that has the lower remaining cost in carrying traffic from origin to destination can clearly afford to charge the lower freight rate for the route and get the business.

Next, suppose that the bridge owner, B, decides to enter the rail-transport business in competition with X and Y. How does that affect the proper price to charge for the bridge? The answer is that all three railroads should now pay the same price. That answer, called the *parity principle*, still leaves a problem, because the bridge owner doesn't really pay a price to itself for use of the bridge. It is still left for the courts and the regulatory agencies to decide what price the bridge owner really is paying for its own railroad's bridge crossings—a price it should also charge its rivals for bridge crossings.

Economists have offered ways of dealing with that problem, too, but the solution they offer is a bit too complicated to describe at this point. (For a quotation of confirmation of the legitimacy of the economists' parity-pricing rule by the judges of the Privy Council in London, see the box below.)

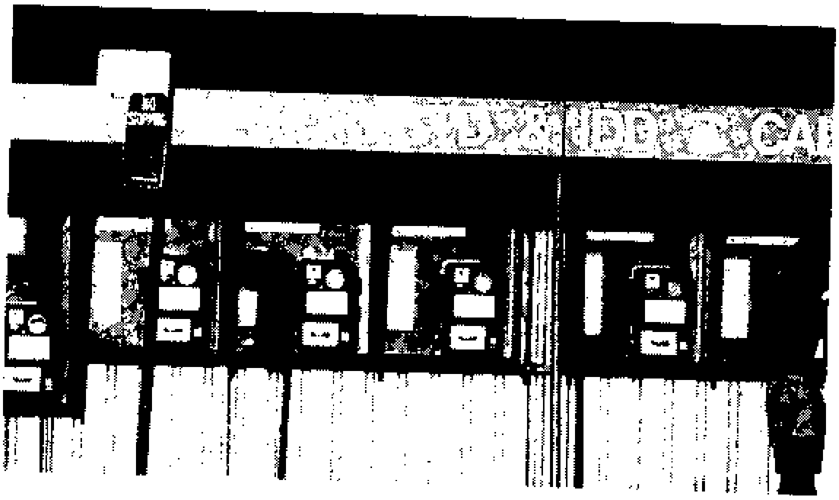
THE PRIVY COUNCIL APPROVES PARITY-PRINCIPLE ACCESS PRICING

New Zealand Telecom, the newly privatized telephone company of New Zealand, soon faced competition from an entrant, Clear Communications. Clear offers long-distance service and hopes to provide local service, as well. For Clear's long-distance messages to reach their intended recipients, it is generally necessary to make use of Telecom's local loop. After negotiations over the price for this bottleneck service broke down, Clear sued Telecom. Telecom offered to adopt the parity-principle charge for access to its local loop, but Clear refused.

The High Court of New Zealand, with some reservations, decided that the parity-principle offer was appropriate, but this decision was overturned by the Court of Appeal on grounds related to special features of New Zealand law. That decision then was reviewed by the final appeals body, the Privy Council in London. In October 1994, the council issued its judgment, fully supporting the parity principle and holding that there were no difficulties in New Zealand law of the sort cited by the Court of Appeal. The decision of the Privy Council stated in part:

> Both the High Court and the Court of Appeal proceeded on the basis, with which their Lordships [that is, the judges of the Privy Council] agree, that if the terms Telecom was seeking to extract were no higher than those which a hypothetical firm would seek in a perfectly contestable market, Telecom was not using its dominant position [that is, it was not behaving as a monopolist]. . . . The [parity principle] rule is a closely reasoned economic model which seeks to show how the hypothetical firm would conduct itself. . . . [T]he underlying object [of the relevant New Zealand law] will be achieved if the [parity principle] rule is applied.

SOURCE: Privy Council Appeal No. 21 of 1994, Judgment of the Lords of the Judicial Committee of the Privy Council, pp. 21 and 27.



SOME EFFECTS OF DEREGULATION

The effects of deregulation in the United States are still being hotly debated. Yet, several conclusions are becoming clear.

1. *Effects on prices.* There seems little doubt that deregulation has generally led to lower prices. Airline fares, railroad freight rates, and telephone rates all declined on the average (in real, inflation-adjusted terms) after deregulation. At least in the case of the airlines, however, the rate of decline slowed abruptly toward the end of the 1980s. Still, observers conclude that most of these prices are well below the levels that would have prevailed under regulation.
2. *Effects on local service.* During the debates on deregulation, it was widely feared, even by supporters of deregulation, that smaller and more isolated communities would be deprived of services because the small numbers of customers would make those services unprofitable. It was said that airlines, railroads, and telephone companies would withdraw from such communities once they were no longer forced to stay there by the regulators. These worries have proved largely groundless. True, the larger airlines have left the smaller communities, as predicted. But they have usually been replaced by smaller commuter airlines, sometimes affiliated with the major airlines, that have provided, on the average, more frequent service than their regulated predecessors. A few communities have been left without service or with service of poorer quality, but other locations have benefited considerably.
3. *Effects on entry.* As a result of deregulation, older airlines invaded one another's routes, and a number of new airlines sprang up. Altogether, some 14 new airlines and about 10,000 new truck operators entered those markets since deregulation. Almost all of the new airlines ran into trouble and were sold to the older airlines. But since 1990, many small airlines have been launched.
4. *Effects on unions.* Deregulation has badly hurt unions such as the Teamsters (of the trucking industry) and the Airline Pilots Association. In the new, competitive climate, firms have been forced to make sharp cuts in their work forces and to resist wage increases and other costly changes in working conditions. Indeed, there has been strong pressure for retrenchment on all of these fronts. It should not be surprising, then, that some of the affected unions have undertaken efforts to get Congress to reimpose regulation.
5. *Effects on concentration and mergers.* Particularly in aviation and rail freight transportation, deregulation was followed by a wave of mergers in which two firms agreed to join together or in which one firm agreed to be bought out by another. This has led to an expansion of the sizes of the largest companies in the affected industries.

That this has happened should not be surprising since, as we saw earlier in the chapter, industries with important economies of scale are the most likely targets for regulation. Once freed from regulatory constraints, it was to be expected that firms in such industries would try to take advantage of opportunities to achieve cost reductions through rapid expansion or by mergers.

Evaluations of the merger movement have differed sharply. Some have concluded that mergers threaten to increase monopoly power and exploit the public. Others have argued that indirect competitive pressures (for example, the rivalry of barges and trucks with large railroads) remain strong and that

economies of scale resulting from the mergers will be passed on to the consuming public.

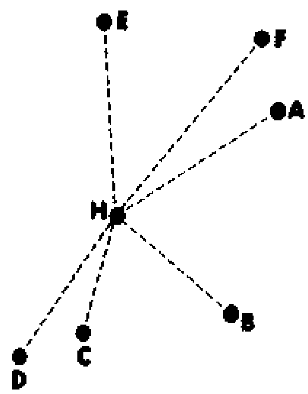
CONSEQUENCES OF DEREGULATION: GENERAL COMMENTS

Deregulation brought few surprises to economists. Reduced prices, reduced costs, increased pressures upon unions, and some rise in mergers were all expected. What *did* come as a surprise was the magnitude of these changes. No one seems to have expected that wages and working hours of pilots employed by the new airlines would differ so sharply from those traditional in the industry. No one seems to have expected that merger activity would be so extensive.

The general public seems to have been unpleasantly surprised in another respect. It was to have been anticipated that increased price competition would bring with it some reduction in "frills." To cut costs in order to reduce prices, airlines have had to make meals less elaborate and less costly. They have had to limit the number of flights to avoid empty seats, and increased crowding of planes is the clear consequence. To fill planes with more passengers, many airlines have turned to "hub and spoke" systems (see Figure 18-2). Instead of running a flight directly from a low-demand airport, A, to another low-demand airport, B, the airline flies all passengers from Airport A to its "hub" at Airport H, where all passengers bound for destination Airport B are asked to board the same airplane. This clearly saves money and gives passengers more options as the number of flights between hubs and spokes increases. But it is not as convenient for air passengers as a direct flight from origin to destination. Critics of deregulation have placed a good deal of emphasis on the reductions in passenger comfort, but economists argue that competition would not bring such results unless passengers as a group prefer the reduction in fares to the greater standards of luxury that preceded them.

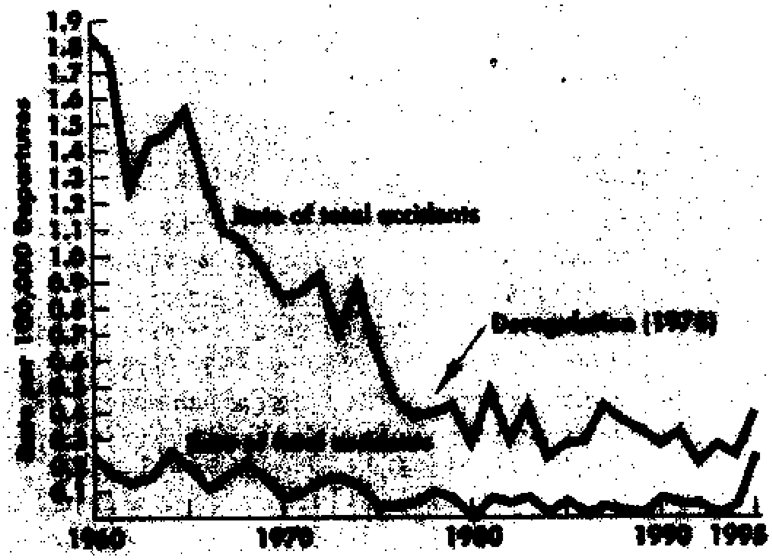
In addition, some observers have been concerned about the safety effects of deregulation. In 1985, when there was an unusually large number of air accidents, critics even implied that this might be attributable to deregulation as airlines cut expenditures on safety to keep prices low. However, in the following year, U.S. commercial airlines achieved an all-time record in terms of passenger safety. Deregulation seems not to have produced any break in the trend toward

FIGURE 18-2 A "HUB AND SPOKE" AIRLINE ROUTING PATTERN



Passengers do not fly directly along the sparsely traveled route from Airport A to Airport B. Instead, passengers from A are flown to the hub airport, H, and then redistributed to an airplane flying to Airport B. Deregulation greatly increased use of this procedure.

FIGURE 18-3 AIRCRAFT ACCIDENTS PER 100,000 DEPARTURES, U.S. SCHEDULED AIRLINES, 1960–1995



Although accident rates vary widely from year to year, both total and fatal accidents per 100,000 departures have declined substantially during this period, both before and after deregulation (which occurred in 1978).

SOURCE: National Transportation Safety Board, various news releases from January 1971 through January 1996.

increased safety in air transportation, as Figure 18-3 shows. Still, deregulation may require special vigilance to guard against neglect of safety as a cost-cutting measure, and the expense of additional government inspection can be considered a required cost of deregulation.

The general conclusion is that deregulation has usually worked out well and promoted the welfare of consumers. But the battle for deregulation is far from over. Even if those who wish for a return to the good old days of regulation (and there are some) do not succeed, there are many areas in which regulation of the old-time variety still continues its grip.

Moreover, there is an ongoing battle between the local telephone companies and cable television firms, each of which wants to invade the other's territory. Because they have wires that enter so many homes, the cable television carriers are virtually the only prospective rivals with the power to create effective competition for the phone companies that supply local service to households and businesses. But similar facts make the telephone firms real competitive threats to cable providers. Predictably, each wants to invade the territory of the other, but to keep the other out of its own. Regulators, with their common predisposition against entry, kept these prospective rivals apart, thereby denying the public whatever benefits their competition may offer. The 1996 Telecommunications Act seeks to eliminate this roadblock.

A WORD ON PRIVATIZATION

As we indicated at the beginning of the chapter, in industries in which monopoly or near monopoly offers cost advantages to society over competition, there is an alternative to regulation. This alternative is government ownership and operation of the firms in that industry, or *nationalization*.

In the United States, tradition does not favor such government operation, but there are exceptions. For example, we have government supply of electricity by

the Tennessee Valley Authority, the U.S. Postal Service, and the operation of passenger railroads by publicly owned Amtrak. A number of cities operate their own public transport facilities, collect their own garbage, generate electricity, and offer other services that elsewhere are provided by private enterprise.

It is almost an instinctive reaction by people in the United States to consider such public enterprises as being prone to extreme mismanagement and waste. The nationalized Swedish telephone system is smooth-working and efficient, and the French government-supplied electricity system has set world standards in its use of the most modern analytic techniques of economics and engineering. Still, in much of Europe and South America there is a major move toward returning nationalized firms to private enterprise, often under regulatory oversight.

It is widely agreed that nationalized industries have weak incentives for efficiency. First, governments virtually never permit a nationalized firm to go bankrupt, and, as a result, management is deprived of one of the most powerful motivations to minimize costs. Second, no one has yet found a systematic incentive mechanism for efficiency that can do for nationalized industries what the profit motive does for private enterprise. Where the market is unsparing in its rewards for accomplishments and in its penalties for poor performance, one can be quite sure that a firm's inefficiency will not readily be tolerated. But nationalized industries have no such automatic mechanism handing out rewards and penalties dependably and impartially.

The conservative government of Margaret Thatcher, by one estimate, "privatized" no less than 40 percent of the British industries that had been nationalized between 1945 and 1979.[1] The list includes telecommunications, oil and gas production, airports and airlines, trucking, some hotels, seaports, ship building, and the aerospace, automobile, and semiconductor industries. Under "Thatcherism," bus routes were also deregulated; local governments began to contract out services to private contractors; private pensions, health care, and education grew; and over a million public housing units were sold to tenants. Television broadcasting, formerly the exclusive province of the BBC, a government-owned corporation, is going increasingly into private hands. The movement has spread to many other geographic areas. For example, privatization of telecommunications has occurred or is contemplated in places as diverse as New Zealand, Venezuela, and Puerto Rico.

Privatization in the free-market economies has not proved to be as easy or trouble-free as some had expected. Many of the government firms that were candidates for sale to the public were such money losers that no one could be found to buy them. In some cases, firms were sold with what amounted to guarantees that they would receive government-enforced monopolies, presumably in order to increase the prices that governments could extract from those sales. In other cases, the firms sold seemed to be natural monopolies; in neither case could the firm's decision making be entrusted to the market mechanism. Consequently, some such companies, British Telecom for example, found themselves enmeshed in regulatory constraints that severely inhibited their decisions. Other such firms, such as New Zealand Telecom, were sold on condition that cross-subsidies (referred to as the "Kiwi share" in New Zealand) and other peculiarities of government operation would be continued under private ownership. Such costly

[1]Martin Holmes, *Thatcherism* (London: Macmillan, 1989); Dennis Swann, *The Retreat of the State: Deregulation and Privatisation in the U.K. and U.S.* (New York: Harvester-Wheatsheaf, 1988); and Christopher Johnson, *The Economy under Mrs. Thatcher* (London: Penguin, 1992).

restrictions predictably lead to expensive litigation between the privatized firms and their rivals over the proper consequences of the obligations adopted as part of the sale and pricing rules that allow no firm an indefensible competitive advantage.

However, it is in the formerly centrally planned economies of eastern Europe that proposed moves away from nationalization sound like a stampede. The transformation of the economies of Poland, Hungary, the Czech Republic, Slovakia, eastern Germany, and Lithuania into free-market systems imply extreme disillusionment with the performance of government-owned industry. (See the box on page 240 in Chapter 10 for more on the Polish transformation.)

INDUSTRY OPPOSITION TO DEREGULATION—THE PUZZLE REVISITED

We now return to the puzzle with which we started this chapter. Why would some members of the trucking industry strongly oppose deregulation of in-state trucking, even though it offered them more freedom? We have already learned that regulation often stifles competition. As it turns out, that lies at the heart of the Teamsters' opposition to deregulation.

Since interstate trucking was deregulated in 1980, at least 40,000 new carriers have entered the industry. In the new, competitive climate, firms have been forced to make sharp cuts in their work forces and to resist wage increases and other costly changes in working conditions. It should not be surprising then that the Teamsters vehemently opposed further deregulation of the industry. The same has been true of unions in other deregulated industries such as airlines. Despite the union's efforts, however, intrastate trucking was deregulated in 1995, and most observers expected the result would be lower prices, better service, and less bureaucratic red tape.

SUMMARY

1. **Regulation** has two primary purposes: to put brakes on the decisions of industries with monopoly power and to contribute to public health and safety.
2. Railroads, trucking, telecommunications, and gas and electricity supply are among the industries that are regulated in the United States. In Europe, the firms that provide these services have, until recently, usually been owned by governments. (They were nationalized.)
3. In recent years there has been a major push toward reduction of regulation. So far, air, truck, and rail transportation have been deregulated in whole or in part.
4. Among the major reasons given for regulation are: (a) **economies of scale** and **economies of scope,** which make industries into natural monopolies; (b) the danger of self-destructive competition in industries with low (short-run) marginal costs; (c) the desire to provide services to isolated areas where supply is expensive and unprofitable; and (d) protection of consumers, employees, and the environment.
5. Regulators often reject proposals by regulated firms to cut their prices, and sometimes the regulators even force firms to raise their prices. The purposes of such actions are to prevent "unfair competition" and to protect customers of some of the firm's products from being forced to **cross-subsidize** customers of other products. Many economists disagree with such actions and argue that the result is usually to stifle competition and make all customers pay more than they otherwise would pay.
6. Economists generally argue that a firm should be permitted to cut its price as long as it covers its *marginal cost*. However, in many regulated industries, firms would go bankrupt if all prices were set equal to marginal costs.
7. Regulation is often criticized for providing little or no incentive for efficiency, for tending to push prices upward, and for forcing the regulated parties to engage in an expensive and time-consuming adversary process.

8. Several regulatory agencies have recently adopted new methods of regulation intended, among other things, to provide incentives for efficiency analogous to those supplied by the free market.
9. To prevent the undermining of competition, the owner of a (monopoly) bottleneck facility must charge all users of the facility (including itself) the same price for access.
10. There has recently been a worldwide movement toward **privatization** of nationalized firms, that is, their sale to private owners.

KEY TERMS

Regulation
Nationalization
Privatization
Price floor
Price ceiling
Natural monopoly
Economies of scale
Economies of scope
Cross-subsidization
Self-destructive competition
Marginal cost pricing
Ramsey pricing rule
Stand-alone cost
Regulatory lag
Price caps
Access price
Bottleneck facility

QUESTIONS FOR REVIEW

1. Why is an electric company in a city often considered to be a natural monopoly? What would happen if two competing electric companies were established? How about telephone companies?
2. Suppose that a 20-percent cut in the price of freight transportation brings in so much new business that it permits a railroad to cut its passenger fares by 2 percent. In your opinion, is this equitable? Is it a good idea or a bad one?
3. In some regulated industries, regulatory agencies prevent prices from falling, and as a result many firms open for business in those industries. In your opinion, is this competitive or anticompetitive? Is it a good idea or a bad one?
4. What industries in the United States can be considered nationalized or partly nationalized? What do you think of the quality of their services? Why might this criterion be inadequate as evidence on which to base a judgment of the idea of nationalization?
5. List some industries with regulated rates whose services you have bought. What do you think of the quality of those services?
6. In which if any of the regulated industries mentioned in your previous answer is there competitive rivalry? Why is regulation appropriate in these cases? (Is it inappropriate in your opinion, and if so, why?)
7. Regulators are much concerned about the prevention of "predatory pricing," that is, pricing policies designed to destroy competition. The U.S. Court of Appeals has, however, noted that "the term probably does not have a well-defined meaning, but it certainly bears a sinister connotation." How might one go about distinguishing "predatory" from "nonpredatory" pricing? What would you do about it?
8. Do you think that it is fair or unfair for rural users of telephone service to be cross-subsidized by other telephone users?
9. Can you think of a way in which a new rural telephone subscriber contributes a beneficial externality? If so, does it make sense to provide a subsidy to rural subscribers, and who should pay the subsidy?
10. Suppose that a firm in a regulated industry is prohibited from earning profits higher than it now is getting, so it begins selling a new product at a price above its long-run marginal cost. Explain why the prices of other company products will, very likely, have to be reduced.
11. To provide incentives for increased efficiency, several regulatory agencies have eliminated ceilings on the profits of regulated firms but instead put caps on their prices. Suppose that a regulated firm manages to cut its prices in half, but in the process, it doubles its profits. Should rational consumers consider this to be a good or a bad development? Why?
12. Why do you think that Great Britain has privatized some of its nationalized industries? How about Poland? Where is privatization more urgent?



CHAPTER 19

Limiting Market Power: Antitrust Policy

Within 10 minutes, 10 men wearing dark business suits emerge from different buildings in downtown San Francisco, and each immediately opens his black umbrella. Isn't that ominous? (Answer: Not if it's raining!)
A law school discussion of suspicious parallel conduct by business firms

The preceding chapter described the process of regulation, one of the two main instruments used by government to offset the undesirable effects of unrestrained monopoly and oligopoly. This chapter analyzes the second of these instruments, antitrust policy. *Antitrust policy* refers to programs that preclude the deliberate creation of monopoly and prevent powerful firms from engaging in related "undesirable practices." Firms accused of violating the antitrust laws are likely to be sued by the federal government or other private firms. These suits are aimed at preventing this undesirable behavior from recurring, providing compensation to the victims, and punishing the offender by fines or even a prison term

Antitrust suits are usually well-publicized affairs because the accused firms are often the giants of industry The more spectacular cases in the history of antitrust policy involve such names as Standard Oil, U.S. Steel, the Aluminum Company of America (Alcoa), General Electric, International Business Machines (IBM), American Telephone and Telegraph (AT&T), and Microsoft Even some of the nation's most prestigious colleges and universities have been accused of engaging in a pricing conspiracy. Recently, the list of antitrust targets has been growing longer, as the Department of Justice appears to have stepped up its enforcement efforts. The number of its antitrust cases increased from 3 in 1992 to 22 in 1994.

The magnitude of an antitrust suit is difficult to imagine. After the charges have been filed, it is not unusual for more than 5 years to elapse before the case even comes to trial. The parties spend this period laboriously and meticulously preparing their cases. Dozens of lawyers, scores of witnesses, and hundreds of researchers are likely to participate in this process. The trial itself is also likely to run for years. A major case can produce literally thousands of volumes of material, and it can easily cost the defendant *several hundred million* dollars.

Thus, the power of the government or another firm to haul a company into court on antitrust charges is an awesome one. The reason is—win, lose, or draw—

such a case constitutes a heavy burden on the accused firm, draining its funds, taking the time and attention of its management, and delaying business decisions until the outcome of the legal proceedings is determined.

What justifies investment of so much power in government agencies such as the Department of Justice or the Federal Trade Commission? What are the purposes of the antitrust laws? How well has the program succeeded? In fact, there is much dispute about whether antitrust laws have done much to increase competition, and some observers have even argued that they are often abused and twisted into anticompetitive tools. (See the box, opposite.) These are the main concerns of this chapter.

Today, a primary issue for U.S. antitrust policy is whether its rules, which are generally more severe than those in other countries, seriously handicap U.S. companies in their efforts to compete in world markets. If so, are the benefits of the antitrust laws sufficient to offset that cost to society? This chapter will provide pertinent evidence. But, since the issue is controversial, we will not attempt categorical answers. The final judgment will be left to you.

THE PUBLIC IMAGE OF BUSINESS WHEN THE ANTITRUST LAWS WERE BORN

The Sherman Act, the forerunner of all modern antitrust legislation, was passed in 1890. To understand what led Congress to interfere with freedom of business enterprise by passing the act, we must glance at the most publicized business practices in the United States during the half-century following the Civil War. No doubt, many businesses were beyond reproach, but these were not the ones that made headlines and yielded spectacular fortunes. Stories of the adventures of the more daring breed of entrepreneurs, those who have been described as "the robber barons," competed in lurid detail with the tales of their contemporaries in the Wild West.

One of the most widely publicized cases was that of John D. Rockefeller, Sr., and his Standard Oil Company. About 5 years after starting in the oil-refining business with an investment of $4,000, Rockefeller and his partners created the Standard Oil Company in 1870. Under its leadership, a number of refineries and other shippers formed a cooperative powerful enough to force the railroads not only to provide discounts to members of the group *and not to its competitors*, but even to give the group "drawbacks"—that is, payments on every shipment of oil refined by a *rival* firm. By 1879, Standard Oil and its associated companies were producing some 90 percent of the nation's refined oil and had control of all of its pipeline capacity.

Then, in 1882, lacking confidence in the trustworthiness of the alliance and because of legal obstacles to its interstate operations, the 40 associated firms formed the Standard Oil Trust (from which the word *antitrust* was derived). The trust closed down "excessive" and inefficient refinery operations, involving more than half of its plants, to limit output, raise prices, and earn monopoly profits.

CAN ANTITRUST LAWS BE USED TO PREVENT COMPETITION?

Many observers are concerned that the antitrust laws are often used by inefficient firms to protect themselves from the competition of more efficient rivals. When they are unable to win out in the marketplace, the argument goes, firms simply file lawsuits against their competitors claiming that those rivals have achieved success by means that violate the antitrust laws.

Not only do they seek protection from the courts against what they describe as "unfair competition" or "predatory practices," but they often sue for compensation which, under the law, can sometimes be three times as large as the damages that they claim to have suffered. Moreover, even if the defendant is found innocent, it must normally pay the very high costs of the litigation itself. Aside from the enormous waste that such suits entail, observers worry that this is a perversion of the antitrust laws, which were, after all, designed to *promote* competition, *not* to *prevent* it.

Two recent examples illustrate the nature of such litigation.* These cases also show that the courts are often wise enough to throw out such attempts to use the antitrust laws to prevent competition.

AMI versus IBM. Allen-Myland Inc. (AMI) is a small firm specializing in the upgrading of computers, an activity which had earned it handsome profits in a period of time when expanding a computer's capacity was very laborious. However, technological progress by IBM then transformed what was once a labor-intensive task requiring considerable skill into the routine installation of a small and highly reliable part. This rendered obsolete many of the services offered by AMI. AMI sued IBM, seeking to persuade the court to impose an artificial and expensive market niche for upgrading services, with AMI permanently protected from competitive pressures. The court's decision completely rejected AMI's position (Eastern District of Pennsylvania [1988]). The decision is under appeal.

Sewell Plastics versus Coca-Cola, Southeastern Container et al. The Sewell Plastics Company once had a preponderant market share in the manufacture of plastic soft-drink bottles in the United States, selling 2-liter bottles for more than 30 cents each. A group of Coca-Cola bottlers in the Southeast considered the price too high and formed a cooperative firm (Southeastern Container) to manufacture their own plastic bottles.

Within 5 years, Southeastern had reduced its cost below 14 cents per bottle, and real retail prices of soft drinks also fell. Despite rising national sales and profits, Sewell sued Southeastern, seeking to force its owners to sell the firm to Sewell or, as a possible alternative, to force Southeastern's customers to sign exclusive purchasing contracts with Sewell. In the spring of 1989, the judge dismissed Sewell's claims, holding that there was no need for a trial (U.S. District Court, Western District, North Carolina [April 1989]).

*One of the authors of this book was involved as an expert witness in both cases.



Other trusts soon followed in sugar, whiskey, lead, and cottonseed and linseed oil. In 1892, the Supreme Court of Ohio ordered the dissolution of the Standard Oil Trust, which nevertheless survived as a cooperating set of firms, with directors of the major refining companies serving on each other's boards.

Other, more lurid tales of business practices in this period are easy to find: J. P. Morgan hired an army of toughs to engage literally in pitched battle for a

contested section of railroad outside Binghamton, New York; Philip Armour and his confederates obtained control of meat processing by an understanding with their rivals that each day a different one of them would offer a low bid for the morning shipment of cattle and no one else would ever enter a higher bid.

Business leaders also repeatedly indicated their contempt for the public interest. J. P. Morgan announced, "I owe the public nothing," and people long remembered W. H. Vanderbilt's phrase "the public be damned." Advocates of control measures warned the public that it faced a country "in which the citizen was born to drink the milk furnished by the milk trust, eat the beef of the beef trust, illuminate his home by grace of the oil trust and die and be carried off by the coffin trust."[1] The circumstances in 1890 were clearly propitious for some legislative action.

THE ANTITRUST LAWS

Five acts of Congress constitute the basis of the federal government's antitrust policy. Major provisions of these acts are summarized in Table 19-1. The *Sherman Act*, the first of the U.S. antitrust laws, was passed in 1890, soon after the trust-creating activity reached its peak. The act is brief and very general, containing two main provisions: It prohibits all contracts, combinations, and conspiracies in restraint of trade (Section 1) and bans monopolization in interstate and foreign trade (Section 2).

However, the Sherman Act provided no definition of its terms and no special agency to oversee its enforcement. Thirteen years elapsed after its passage before the antitrust division of the Department of Justice was established under the energetic antitrust proclivities of Theodore Roosevelt.

During Woodrow Wilson's administration, many believed that the Sherman Act did not provide adequate protection to the public. Consequently, in 1914, Congress passed two supplemental laws, the Clayton Act and the Federal Trade Commission Act.

The *Clayton Act* deals with certain specific practices thought to be conducive to the encroachment of monopoly. First, it prohibits *price discrimination* (see Chapter 11), which it defines as the act, by a seller, of charging different prices to different buyers of the same product. This provision would, for example, have banned the railroad rebates that Rockefeller had used to squeeze out his rivals.

Second, the Clayton Act prohibits *tying contracts*—arrangements under which a customer who wants to buy some product from a given seller is required as part of the price to agree to buy some other product or products exclusively from that same seller.

Third, the Clayton Act prohibits one firm from purchasing the stock of another if that acquisition tends to reduce competition. While this provision was intended to prevent a firm from buying out its rivals, businesses found it possible to circumvent the intent of the law by buying a rival's stocks and then merging assets. When this practice was recognized, a new law—the *Celler-Kefauver Antimerger Act* of 1950—was enacted to prohibit it.

Finally, the Clayton Act prohibits *interlocking directorates* between competitors. Under this arrangement, two companies have in common some of the members of their boards of directors.

---

[1]Matthew Josephson, *The Robber Barons, The Great American Capitalists, 1861–1901* (New York: Harcourt Brace, 1934), p. 358.

TABLE 19-1 BASIC ANTITRUST LAWS

Name	Date	Major Provisions
Sherman Act	1890	Prohibits "all contracts, combinations and conspiracies in restraint of trade" (Section 1) and monopolization in interstate and foreign trade (Section 2).
Clayton Act	1914	Prohibits price discrimination; contracts in which sellers prevent buyers from purchasing goods from the sellers' competitors (tying contracts); and acquisition by one corporation of another's shares if these acts are likely to reduce competition or tend to create monopoly. Also prohibits directors of one company from sitting on the board of a competitor's company.
Federal Trade Commission Act	1914	Established the FTC as an independent agency with authority to prosecute unfair competition and to prevent false and misleading advertising.
Robinson-Patman Act	1936	Prohibits special discounts and other discriminatory concessions to large purchasers unless based on differences in cost or "offered in good faith to meet an equally low price of a competitor."
Celler-Kefauver Antimerger Act	1950	Prohibits any corporation from acquiring the assets of another where the effect is to reduce competition substantially or to tend to create a monopoly.

The *Federal Trade Commission Act* created a commission to investigate "unfair" and "predatory" competitive practices and declared illegal all "unfair methods of competition and commerce." But since the law provided no definition of "unfairness," and since, in any event, the commission's powers were substantially restricted by the courts, the FTC was a rather ineffective agency for the first quarter-century of its existence. In 1938, however, it was given the task of preventing false and deceptive advertising, a task to which it has subsequently devoted a substantial portion of its energies.

In 1936, Congress passed the *Robinson-Patman Act* to protect independent sellers—both wholesalers and retailers (primarily in groceries and drugs)—from the "unfair competition" of chain stores and mass distributors. The Robinson-Patman Act was not a natural step in the succession of antitrust laws, since it sought to *restrain* competition by protecting small firms from the competition of larger ones. It was believed that large firms were powerful enough to wrest special financial terms from their suppliers, which gave them an unfair competitive edge over their rivals. Accordingly, the act prohibited several types of discriminatory arrangements, including special concessions like promotional allowances by sellers to any favored set of buyers, along with discounts for larger purchases or any other form of discrimination that tended to *reduce* competition or encourage monopoly.

THE COURTS AND THE SHERMAN ACT

From the earliest days of the Sherman Act, the courts have been rather consistent in their use of Section 1—the part of the act that prohibits all contracts, combinations, and conspiracies in restraint of trade. Section 1 has been invoked primarily against price-fixing agreements—that is, agreements under which several ostensibly competing firms coordinate their pricing decisions. The courts have held that such agreements are illegal per se; that is, they have held that no excuses or exonerating circumstances can render a price-fixing agreement acceptable under the law.

This doctrine has been invoked many times, most notably in the General Electric–Westinghouse case decided in 1961. General Electric, Westinghouse, and

PROTECTION OF COMPETITION, NOT PROTECTION OF COMPETITORS

The courts have repeatedly emphasized that the antitrust laws are not intended to make life easier for individual firms that encounter difficulties coping with competitive market pressures. The following quotation from a recent decision of the U.S. Supreme Court makes this clear:

> The purpose of the Sherman Act is not to protect businesses from the working of the market, it is to protect the public from the failure of the market (*Spectrum Sports Inc. v. McQuillan*, US 122 L.Ed 2d 247 [1993], p. 506).

Another recent decision by the U.S. Court of Appeals, 7th Circuit, indicates why firms do not like the market forces:

> Competition is ruthless, unprincipled, uncharitable, unforgiving—and a boon to society, Adam Smith reminds us, precisely because of these qualities that make it a bane to other producers (*Stamatakis Industries v. King*, 965 F 2d 469, 1992, p. 471).

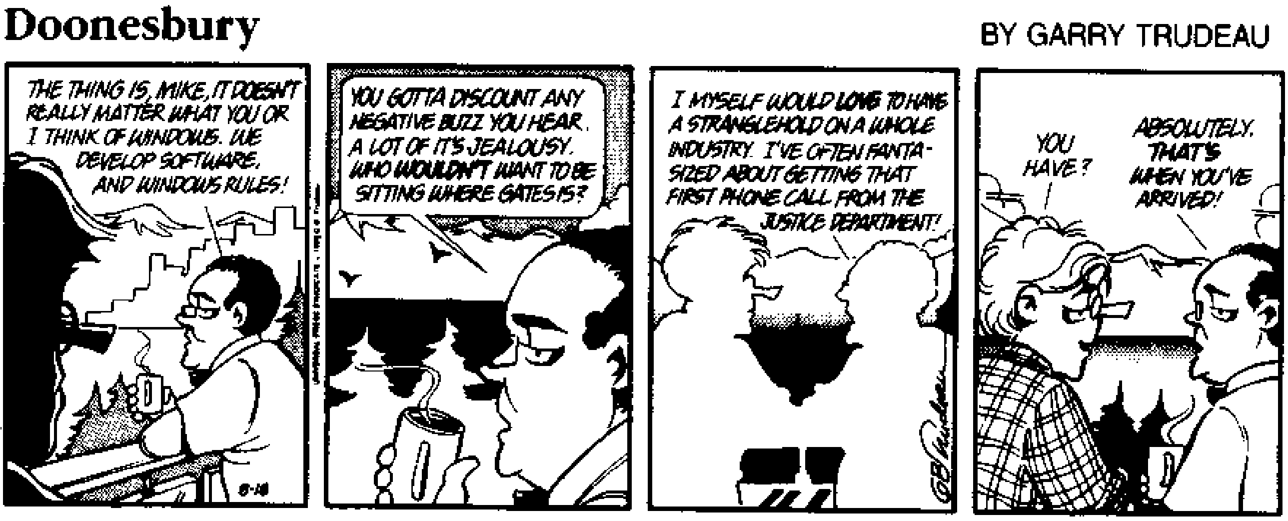
several dozen other producers of electrical equipment had gotten together to agree on prices and divide up the market among themselves. The firms were found guilty of a conspiracy to fix prices and were fined several million dollars. Even more remarkable, officers of the major companies were sentenced to (brief) prison terms.

Section 2 of the Sherman Act deals with individual decision makers, rather than with groups of conspirators, focusing on attempts to create monopolies. At first, the Supreme Court proceeded timidly in dealing with industrial cases under Section 2. For example, an 1895 ruling held that a monopoly of sugar manufacturing was legal on the grounds that manufacturing was not commerce! But the court's position toughened markedly in 1911, when it decided to require both the American Tobacco Company and the Standard Oil Company to give up substantial shares of their holdings in other firms. Many of today's leading gasoline suppliers—including Standard Oil of California, Exxon, and Sohio—are offspring of the original Standard Oil Company, spawned by the court's decision.

At the same time, however, the court also formulated the troublesome *rule of reason*, which held that monopolizing trade restraints are not *necessarily* illegal per se. According to this rule, a restraint is against the law only if it is "unreasonable." The court held (and many economists agree) that size, by itself, does not constitute an offense—that a firm must commit objectionable overt acts before it can be found guilty of violating Section 2 of the Sherman Act. Eastman Kodak and International Harvester, each with very large market shares, were found not guilty on similar grounds. More recently, charges against IBM and Microsoft were not pursued or were settled by compromise, despite the large relative sizes of these firms. Thus, while the courts held that there were *no* excusable cases of price fixing under Section 1, they ruled that there *were* excusable monopolies under Section 2.

TACIT COLLUSION

As just noted, Section 1 of the Sherman Act prohibits explicit agreements among direct competitors on the prices they will charge. Most firms have learned to avoid discussions of price with their rivals. As a result, when several firms are accused of conspiring, the collusion is usually claimed to be "tacit." Firms are alleged to have signaled to one another. For example, one may announce to the press that it is planning a price increase, hoping that competitors will take the hint and raise their prices as well.



Antitrust Reaches the Comic Strips: In this Doonesbury strip the joke is that a U.S. firm reaches true importance only after being investigated by the antitrust division of the U.S. Department of Justice. Bill Gates, referred to in the strip, is the founder and chairman of Microsoft, the world's largest software company, and the creator of the Windows program. The company has been investigated and sued on antitrust grounds a number of times.

Evidence to back up such a claim is not easy to provide, however. First, if the industry consists of a small number of firms of substantial size, each firm clearly recognizes that its decisions are *interdependent* with those of its rivals. Firm A knows that if it starts a large advertising campaign, Firm B will probably respond with a campaign of its own or with some other countermeasure. A also knows that if it raises its price and B does not follow, then A will probably have to withdraw its increase, with loss of customer goodwill and other forms of damage. This happens regularly in the airline industry.[2] Thus, it is often difficult to determine when this crosses into something resembling collusion.

Second, if rival firms sell virtually identical products, market forces do not permit one firm to charge a price much higher than another's. Consequently, all firms in the industry will ultimately end up charging approximately the same price—not because they want to, but because market forces give them no choice. Nonetheless, this has not stopped critics from leveling collusion charges against certain companies. For example, in 1995 the big four cereal makers were accused of colluding to keep prices artificially high. Kellogg, General Mills, Post, and Quaker Oats, which together control 85 percent of the U.S. cereal market, have denied the allegations.

When does parallel conduct constitute evidence of collusion? This question is what the opening quotation of the chapter is all about. It may seem rather strange and ominous if 10 men emerge from 10 doorways on different streets and all simultaneously open their umbrellas—unless it is raining. Rain provides a logical explanation for behavior that is clearly innocent and coincidental!

Thus, one type of evidence of "tacit collusion" that does seem to stand up in court is observed behavior that makes sense if the firms are engaged in a

[2]But firms that recognize their interdependence must surely take it into account in making decisions. It would be irrational and irresponsible of them not to do so. Nor is it illegal for a firm to take its rival's likely reactions into account. But when does this cross the borderline into something resembling "collusion"?

conspiracy but makes no sense otherwise. Here is an example. Suppose that all seven firms in an industry have clearly inelastic demands and are losing money. Then six of them suddenly reduce their prices sharply. Is this valid evidence of a conspiracy to destroy the seventh firm? Very possibly, because, as we know, a cut in price can reduce losses only when demands are elastic. The only sensible explanation of the behavior of the six firms seems to be that they were colluding to reduce competition.

PREDATORY PRICING

PREDATORY PRICING is pricing that threatens to keep a competitor out of the market. It is a price that is so low that it will be profitable for the firm that adopts it only if a rival is driven from the market.

A **MERGER** occurs when two previously independent firms are combined under a single owner or group of owners. A *horizontal merger* is the merger of two firms producing similar products, as when one toothpaste manufacturing firm purchases another. A *vertical merger* involves the joining of two firms, one of which supplies an ingredient of the other's product, as when an automaker acquires a tire manufacturing firm. A *conglomerate merger* is a union of two unrelated firms, as when a defense contractor joins a firm that produces videotapes.

One fairly common accusation in antitrust cases is the claim that the defendant has adopted unjustifiably low prices, either to drive a competitor out of business or to prevent the entry of a rival. This practice is called **predatory pricing.** Deciding whether pricing is "predatory" is difficult, both for economists and for the courts, because low prices generally benefit consumers. Therefore, the courts should not want to discourage firms from cutting prices by being too eager to declare that lower prices are intended to destroy a rival.

One principle widely followed by the courts holds that prices are predatory only if they are below either marginal or average variable costs. The logic of this criterion as a test for whether prices are "too low" is that even under perfect competition, prices will not, in the long run, fall below that level, but will equal marginal costs. Even in cases where prices are below marginal or average variable costs, they may be held to be predatory only under two conditions: (a) if there is evidence that the low price threatened to destroy a rival or keep it out of the market and (b) there is a real probability that the allegedly predatory firm could raise prices to monopoly levels after the rival was driven out, thereby profiting from its venture in crime.

Many successful firms—including AT&T, Honeywell, and American Airlines—have been accused of predatory pricing, often in private antitrust suits in which the plaintiff was a competitor rather than the Department of Justice or another government agency. The defendants typically argue that their low prices cover both marginal and average variable costs, that the prices are low because of superior efficiency, and that the lawsuit was brought by their competitors in order to prevent the defendants from competing effectively. (See pages 453 and 467 where the use of antitrust action to impede competition is discussed.) The courts have generally accepted these arguments. There have been many predatory pricing cases, but few convictions.

One notable exception was a case brought against Wal-Mart in 1993 by three Arkansas druggists, who argued that the nation's largest retailer was deliberately pricing products below cost to drive them out of business. Although an Arkansas state court ruled that Wal-Mart was guilty of predatory pricing, the ruling was later reversed by a higher court.

MERGER POLICY

Mergers have long been a subject of suspicion by the antitrust authorities. This is particularly true when a merger is *horizontal*, meaning that the merging firms compete directly by supplying products that are identical or very similar. In these cases, it is often feared that competition will decline because there will be fewer firms in the industry (that is, concentration will be increased).

YOU ARE THERE AN ANTITRUST TRIAL

The charming courtroom is old but recently refurbished, and the air conditioning is inadequate. It is often difficult to hear what is happening. The defendant firm has been accused of predatory pricing, that is, of charging very low prices in order to drive a competitor out of the market (see page 458), and is defending itself against a judgment that could run into billions of dollars.

For the past 2 months, both sides have called up many witnesses—company executives, accountants, statisticians. The female lawyers are dressed in conservative outfits, the men in somewhat seedy two-piece matching suits. It would not do to appear too wealthy, for this is a jury trial, and the nine men and women wear casual attire including sneakers, jeans, and sports clothes. While determined to see justice done, they are having a hard time staying awake under a hurricane of technical arguments and contradictory figures.

The judge follows the proceedings closely, often interrupting with questions of her own. Sometimes she jokes with the witnesses.

The lawyers call in an expert witness who is a specialist in the field—in this case an economist who has written on predatory pricing. He explains to the court and the jury the current thinking of the economics profession on the definition of predatory pricing and the standards by which one judges whether or not it has occurred. He is persuasive.

But the judge and jury have already heard from another economist, equally distinguished, representing the other side. Their analyses, which were quite technical, reached opposite conclusions. Which one are the jurors to believe, and on what basis?



The Department of Justice and the Federal Trade Commission are both concerned with mergers. They do not wish to impede mergers that seem likely to increase efficiency by, for example, improving the coordination of production activities, permitting economies of scale, or getting one of the firms out of financial difficulties. But the antitrust agencies do want to prevent mergers that threaten to reduce competition.

To help firms decide whether a proposed merger will get them into trouble, and for other reasons as well, the Department of Justice issues guidelines indicating when it is (or is not) likely to try to block a merger. For example, the department generally will not oppose mergers in industries that are very unconcentrated or into which entry is very easy. However, in highly concentrated industries where entry is difficult, the merger of two large firms will usually be opposed.

During the early 1980s, the Justice Department issued several sets of new, more permissive guidelines. This loosening of the antimerger rules, coupled with the Reagan administration's view that mergers were generally good for the economy, spawned a much-publicized rise in merger activity. In the entire decade from 1972 to 1982, companies spent about $340 billion (in dollars of 1989 purchasing power) buying other firms. In 1988 alone, the total spent on "dealmaking" activity (including mergers, acquisitions, and leveraged buyouts) topped

$330 billion. In total, about $1.5 trillion was spent during the decade of the 1980s.[3]

After a brief slowdown in the early 1990s, the merger pace has picked up again. 1995's record-breaking $458 billion worth of transactions included the merger of Chase Manhattan and Chemical Bank, Disney's acquisition of ABC Capital Cities, and Time Warner's purchase of Turner Broadcasting.

This rash of mergers has given rise to much controversy. Some observers believe that it increases the likelihood of monopoly power, while others believe that it serves largely to make the merged firms more efficient. In any event, defenders argue that mergers have not historically increased the market share of big business in the United States. (See the discussion of concentration in the next section.) For example, in 1994 the share of the labor force employed by corporations listed in the Fortune 500 was 15 percent, barely changed from its 17 percent share in 1970.[4]

Though by no means unanimous on the subject, most economists agree that mergers *sometimes* reduce competition, particularly in markets that are not contestable, so that threats of entry do not prevent the merged firms from raising prices above competitive levels.[5] This danger is particularly acute if the number of firms is small enough to make collusion a real possibility.

On the other hand, where there is reason to believe that mergers will not reduce competition, many economists oppose impediments to them. Economists believe that mergers that are not undertaken to reduce competition can have only one purpose—to achieve greater efficiency. For example, the larger firm that results from the merger may enjoy substantial economies of scale not available to smaller firms. Or, the two merging companies may learn special skills from one another or offset one another's risks.

Mergers have sometimes proved disappointing and brought limited cost savings; a number have subsequently been dissolved. In fact, several studies have found that more than one-third of acquisitions are eventually divested. But economists who defend freedom to merge when there is no demonstrated threat to competition pose a challenging question: Who can judge better than the firms involved whether their marriage is likely to make their activities more efficient? Indeed, a recent study of roughly 22,000 large manufacturing establishments, of which 1,100 had been purchased and merged ("taken over") between 1981 and 1986, found that the merged manufacturing plants subsequently had rates of productivity growth that were some 14 percent higher than other firms in the same industries.[6]

■ ISSUES IN CONCENTRATION OF INDUSTRY

Having reviewed the antitrust laws and their interpretation by the courts, the next logical question is: Do they work? One rough way to measure the success of antitrust legislation is to look at what has happened to the share of American

[3] *Business Week*, January 15, 1990, pp. 52–53, which cites M&A Data Base, *Mergers and Acquisitions*. These figures are all expressed in dollars of 1989 purchasing power

[4] *Fortune*. May 1971 and May 15, 1995.

[5] See Chapter 12, pages 295–296, for a definition and discussion of contestable markets.

[6] Frank Lichtenberg and David Siegel, *The Effects of Leveraged Buyouts on Productivity and Related Aspects of Firm Behavior*. NBER Working Paper No. 3022, 1989.

business in the hands of the largest firms. Some observers, including Marxists, have asserted that one of the basic tendencies of capitalism is toward *concentration of industry*, because small firms are increasingly driven out of business, especially during economic crises, and large firms consequently acquire ever-larger shares of the market.

To test this hypothesis, one can investigate whether there has been such a tendency observed in the United States. If, in fact, concentration has *not* increased, someone who holds these views might be led to surmise that the antitrust program has had a hand in preventing the growth of monopoly. But first we should consider what might have been expected to happen to concentration in the United States in the absence of any countermeasures by government. Is there good reason to expect an inexorable trend toward bigness?

There are two basic reasons why the larger firms in an industry may triumph over the small ones. First, larger firms may obtain monopoly power, which they can use to their advantage. They can force sellers of equipment, raw materials, and other inputs to give them better terms than are available to small competitors, and they can also force retailers to give preferences to their products. These are, of course, the sorts of advantages to bigness that the antitrust laws were designed to eliminate.

The second reason why an industry's output may tend to be divided among fewer and larger firms with the passage of time has to do with technology. In some industries, fairly small firms can produce as cheaply as large ones or even more cheaply, while in other industries only rather large firms can achieve maximal economy. By and large, the difference in number of firms from one industry to another has tended to correspond to the size of the firm that is least costly. Automobile and airplane manufacturing are industries in which tiny companies cannot hope to produce economically; indeed, these are industries made up of relatively few, large firms. In clothing production and farming, matters go quite the other way.

Frequently, innovation seems to have increased the plant size that minimizes costs. Such examples as automated processes or assembly lines suggest that new techniques always call for gigantic equipment, but this is not always true. For example, the invention of truck transportation took much of the freight-shipping market away from the giant railroads and gave it to much smaller trucking firms. Technological change also seems to have favored the establishment of small electronics firms. Similarly, the continued development of cheaper and smaller computers is likely to provide a competitive advantage to smaller firms in many other industries. Furthermore:

> If innovation provides increased cost advantages to larger firms, the growth of firms will be stimulated. However, a fall in the number of firms in the industry need not inevitably result. If demand for the industry's output grows faster than the optimal size of firms, we may end up with a larger number of firms, each of them bigger than before, but each having a smaller share of an expanded market.

For example, suppose that a new process is invented that requires a far larger scale of operation than currently is typical in some industry. Specifically, suppose that the least costly plant size becomes twice as large. If demand for the industry's product increases only a little, we can expect a decrease in the number of firms. But if, because the new process reduces costs or improves the product significantly, the quantity of the industry's product demanded triples at the same time, then the optimal number of firms will in fact increase to one and

one-half times the original number. Each firm will be twice as big as before, so that together they will serve three times the volume. In such a case, each firm's share of industry output will in fact have declined.

In the 19th century, technological developments did seem to call, predominantly, for larger firms that could take advantage of economies of scale. If the same has been true in the current century, and if this trend has somewhat outstripped even the rate of growth in output—that is, the growth of GDP—we should expect some fall in the number of firms in a typical industry. However, as was just noted, not all technological change works in this direction. For example, many firms in the electronics and software industries are relatively small, and there are observers who argue that new techniques will permit smaller firms to supply some telecommunications services without incurring high costs. We must turn to the evidence to judge whether or not American industry has grown more concentrated.

EVIDENCE ON CONCENTRATION IN INDUSTRY

A **CONCENTRATION RATIO** is the percentage of an industry's output produced by its *four* largest firms. It is intended to measure the degree to which the industry is dominated by large firms

There have been many statistical studies of concentration in American industry. One common way of measuring concentration is to calculate the share of the industry's output produced by the four largest firms in an industry, the so-called **concentration ratio.** Of course, there is no reason why the three or five or ten largest firms should not be used for the purpose, but four firms are the conventional standard.

Table 19-2 shows concentration ratios in a number of industries in the United States. We see that concentration varies greatly from industry to industry: automobiles; electric lamps, bulbs, and tubes; and breakfast cereals are produced by

TABLE 19-2 1992 CONCENTRATION RATIOS FOR REPRESENTATIVE INDUSTRIES (PERCENT OF OUTPUT PRODUCED BY THE FOUR LARGEST FIRMS IN EACH INDUSTRY)

Industry	Four-Firm Ratio	Industry	Four-Firm Ratio
Electric lamps, bulbs, and tubes	86	Bottled and canned soft drinks	37
Cereal breakfast foods	85	Motors and generators	36
Motor vehicles and car bodies	84	Dolls and stuffed toys	34
Hard surface floor coverage	83	Boat building and repairing	32
Aircraft	79	Pharmaceutical preparations	26
Rubber tires and inner tubes	70	Musical instruments	25
Primary aluminum	59	Brooms and brushes	23
Fabricated metal cans	56	Fluid milk	22
Ship building and repairing	53	Bolts, nuts, rivets, and washers	17
Fasteners, buttons, needles, and pins	41	Jewelry, precious metal	16
Prerecorded records and recorded tapes	40	Apparel: women's, misses', juniors' dresses	11
Apparel: men's and boys' suits and coats	39		

SOURCE: U.S. Bureau of the Census, "Concentration Ratios in Manufacturing," *1992 Census of Manufactures*. Thanks to Andrew W. Hait, Special Reports Branch, U.S. Bureau of the Census, for the early release of this data, January 1996.

TABLE 19-3 THE TREND IN CONCENTRATION IN MANUFACTURING INDUSTRIES (1901–1992)

	Around 1901	1947	1954	1958	1963	1966	1970	1972	1982	1987	1992
Percentage of value added in industries with four-firm concentration ratios over 50 percent	32.9	24.4	29.9	30.2	33.1	28.6	26.3	29.0	25.2	27.9	26.4

SOURCES: P. W. McCracken and T. G. Moore, "Competition and Market Concentration in the American Economy," Subcommittee on Antitrust and Monopoly, U.S. Senate, March 29, 1973; F. M. Scherer, *Industrial Market Structure and Economic Performance* (Boston: Houghton Mifflin, 1980), p. 68; F. M. Scherer and David Ross, *Industrial Market Structure and Economic Performance*, 3rd ed. (Boston: Houghton Mifflin, 1990), p. 84; personal communication with Professor F. M. Scherer, March 10, 1993; and personal communication with Andrew W. Hait, Special Reports Branch, U.S. Bureau of the Census, January 18, 1996.

highly concentrated industries, while the jewelry, clothing, and soft-drink industries show very little concentration.

Since the beginning of this century, concentration ratios in the United States have, on the average, remained remarkably constant. It has been estimated that, at the turn of the century, 32.9 percent of manufactured goods were produced by industries in which the concentration ratios were 50 percent or more (meaning that at least 50 percent of industry output was produced by the four largest firms). By 1963, the figure had risen only to 33.1 percent, and by 1992, it had actually fallen to 26.4 percent. These figures and those for other years are shown in Table 19-3. Over the course of the 20th century, concentration in individual U.S. industries has shown no tendency to increase.

Since concentration is intended as a measure of the "bigness" of firms in an industry, such information may suggest that the antitrust laws have to some degree been effective in inhibiting whatever trend toward bigness may in fact exist. But even this very cautious conclusion has been questioned by some observers, who argue that these laws have made virtually no difference in the size and the behavior of American business. Is it desirable for antitrust, or other such programs, to inhibit concentration or bigness? That is the issue to which we turn next.

THE PROS AND CONS OF BIGNESS

Why has antitrust regulation become so accepted a part of government policy? Are the effects of bigness or monopoly always undesirable? We know that monopoly power can be abused; the history of the Rockefellers, the Armours, and the Morgans described at the beginning of this chapter confirms that fact. But even when the giants of business are not so swashbuckling in their operations, unrestrained monopoly and bigness give rise to a number of problems:

1. *Distribution of income.* The flow of wealth to firms with market power—and thus to those that can influence prices in their favor—is widely considered to be unfair and socially unacceptable.
2. *Restriction of output.* We learned in Chapter 11 that an unrestrained monopoly maximizes profits by restricting its output below the amount that would be provided by an equivalent competitive industry. This means that

unregulated, monopolized industries are likely to produce smaller outputs than the quantities that best serve society's interests.

3. *Lack of inducement for innovation.* It is sometimes argued that firms in industries with little or no competition are under less pressure to introduce new products and production methods than firms in industries in which each is constantly trying to beat out the others. Without competition, the management of a firm may choose the quiet life, taking no chances on risky investments in research and development. But a firm that operates in constant fear that its rivals will come up with a better idea, and come up with it first, can afford no such luxury.

So far we have presented only one side of the picture. In fact, bigness in industry need not be advantageous only to the firm. It can also, at least *sometimes*, work to the advantage of the general public. Again, there are several reasons:

1. *Economies of large size.* Probably the most important advantage of bigness is to be found in industries where technology dictates that small-scale operation is inefficient. One can hardly imagine the costs if automobiles were produced in little workshops rather than giant factories. The notion of a small firm operating a long-distance railroad does not even make sense, and a multiplicity of firms replicating the same railroad service would clearly be incredibly wasteful.

   On these grounds, most policymakers have never even considered any attempt to eliminate bigness. Their objective, rather, is to curb its potential abuses and, at the same time, try to help the public benefit from its advantages. Of course, it does not follow that every industry in which firms happen to be big is one in which big firms are best. There are observers who argue that many firms, in fact, exceed the sizes that are required for cost minimization.

2. *Required scale for innovation.* Some economists have argued that only large firms have the resources and the motivation for really significant innovation. While many inventions are still contributed by individuals, to put a new invention into commercial production is often an expensive, complex venture that can be carried out only on a large scale. Only large firms can afford the funds and bear the risks that such an effort demands. In addition, according to this view, only a large firm has the motivation to lay out the funds required for the innovation process since it gets to keep a considerable share of the benefits. A small company, on the other hand, will find that its innovative idea is soon likely to be followed by close imitations, which enable competitors to profit from its research outlays.

There have been many studies of the relationship between firm size, industry competitiveness, and the level of expenditure on research and development (R&D). While the evidence is far from conclusive, it does indicate that highly competitive industries composed of very small firms tend not to spend a great deal on research. Up to a point, R&D outlays and innovation seem to increase with size of firm and concentration of industry. However, some of the most significant innovations introduced in this century have been contributed by firms that started very small. Examples include the electric light, alternating current electricity, the photocopier, the electronic calculator, and the desktop computer.

Other Government Programs Related to Bigness

Because the issues raised by bigness and concentration are complex, they call for a variety of policy measures. Certainly, antitrust programs alone cannot do everything that the public interest requires. For example, in cases where one large firm is far more efficient than several small ones—that is, where the industry is a natural monopoly—it does not seem reasonable to break up the industrial giants. Instead, the monopoly firm is *regulated* in one of the ways described in the previous chapter.

A **patent** is a temporary grant of monopoly rights over an innovation.

The concern that competition may inhibit innovation is another important issue. The main instrument government has employed in this area is the **patent** system, which rewards the innovator by granting a temporary monopoly. The patent restricts imitation and is designed to offer small-firm innovators the same advantages from their research activities as are enjoyed by innovators in industries that contain no competitors. Thus, somewhat ironically, while government prohibits monopolization, it also guarantees monopoly power to protect innovative firms in competitive industries. Of course, sometimes the protected firms themselves grow big with the help of the protection. Once-small firms like Polaroid and Xerox grew into industrial giants with the help of government protection through the patent laws.

Questions have been raised about whether patents are effective in inducing expenditures on R&D, and the evidence certainly does not provide overwhelming support for the view that patents strongly stimulate innovation. There also is some debate about the desirability of granting an innovator an unrestricted monopoly for 17 years, as the U.S. patent program now does. Similar issues have been raised about copyright laws, which restrict reproduction of written works.

Finally, government has provided special help to small business in a variety of ways. For example, there are programs designed to make it easier for small firms to raise capital, and special government agencies, such as the Small Business Administration, have been set up for that purpose. There is also some degree of *progressivity* in business taxation, meaning that smaller firms are subject to lower tax rates than those paid by larger firms.

Issues in Antitrust Policy

In recent decades, there has been a searching reexamination of government policy toward business. There have not only been repeated calls for a reduction in the powers of the regulatory agencies but, as we saw in the last chapter, considerable movement in that direction. Some critics have called for abolition of the antitrust laws altogether, while others have advocated their strengthening and expansion. But even if one grants the desirability of an antitrust program with teeth, questions still remain about who or what it should bite.

Structure versus Conduct

One major issue is the relative weight that should be assigned to *structure* and *conduct* in deciding which firms it is in the social interest to prosecute. Most people agree that price fixing or threats of physical violence are socially damaging and should be discouraged. But there is often disagreement over what other types of conduct are undesirable.

Many more questions are raised about the use of structural criteria in antitrust policy. Is bigness always undesirable per se? What if the large firm is more efficient and has engaged in no practices that can reasonably be considered to constitute predatory competition? Many economists have reservations about the prosecution of such a firm, fearing that such a move will only grant protection to inefficient competitors at the expense of consumers. They also point out the danger that successful firms will be singled out for attention under the antitrust laws simply because their success makes them noticeable and their efficiency enables them to outstrip their competitors. The fear is that such an orientation will discourage efficiency and entrepreneurship and reduce competition.

CONCENTRATION AND MARKET POWER

MARKET POWER is the ability of a firm to raise its price significantly above the competitive price level and to maintain this high price profitably for a considerable period.

In this chapter, as in many other discussions of antitrust issues, much was said about concentration. Why should anyone care about concentration ratios? One should care about them if they are a good measure of **market power.** Market power is usually defined as the ability of a firm to raise its price significantly above the competitive price level and to maintain this high price profitably for a considerable period. The question, then, is this: If an industry becomes more concentrated, will the firms necessarily increase their ability to institute profitable increases in prices above competitive levels?

Many economists have concluded that this does not necessarily happen. Specifically, the following three conclusions are now widely accepted:

1. If, after an increase in concentration, an industry still has a very low concentration ratio, then its firms are very unlikely to have any market power either before or after the rise in concentration.
2. If circumstances in the industry are in other respects favorable for successful price collusion (tacit or explicit agreement on price), a rise in concentration will facilitate market power. It will do so by reducing the number of firms that need to be consulted in arriving at an agreement and by decreasing the number of firms that have to be watched to make sure they do not betray the collusive agreement.
3. Where entry into and exit from the industry are easy and quite costless, that is, where the market is highly *contestable*, then even when concentration increases, market power will not be enhanced because an excessive price will attract new entrants who will soon force the price down.

COOPERATION IN RESEARCH BY COMPETING FIRMS

Innovation, and the research that must underlie it, has grown increasingly expensive. Firms are therefore frequently inclined to work cooperatively with other companies in the same industry in financing, planning, and conducting research activities, with all of the participants sharing in the end product. America's antitrust laws have been said to pose a threat to such cooperative activity by competing firms. In no other country is such cooperative activity subject to as great a danger from the law.

The American antitrust authorities have stated repeatedly that they do not want to interfere with research and development; they have, in fact, largely avoided any steps that discourage it. Still, the threat is there, and it may well have inhibited some of the cooperative research activity that U.S. firms might otherwise have undertaken. Thus, there is good reason to consider whether explicit amendments to U.S. antitrust laws—so far as they relate to cooperation

in research—might not contribute to productivity growth in American industry and to its ability to compete effectively in the international marketplace.

USE OF ANTITRUST LAWS TO PREVENT COMPETITION

Finally, let us turn to an important issue that we mentioned at the beginning of the chapter: the misuse of the antitrust laws to prevent competition. There is no doubt that many firms that have been unable to compete effectively on their own merits have turned to the courts to seek protection from successful competitors. They have sometimes succeeded.

Those who try to protect themselves in this way always claim that their rivals have not achieved success through superior ability but, rather, by means that are described as "monopolization." Sometimes the evidence is clear-cut, and the courts can readily discern whether the accused firms have violated the antitrust laws or whether they have simply been too efficient and innovative for the complaining competitors' tastes. In other cases, however, the issues are complicated, and only a long and painstaking legal proceeding offers any prospect of resolving them.

This sort of litigation is almost unheard of in Japan, which is alleged by some critics to offer a great advantage to Japanese firms in the international marketplace. Various steps have been suggested to deal with the misuse of U.S. antitrust laws. In one proposal, if the courts decide that a firm has been falsely accused by another of violating the antitrust laws, then (as is done in other countries) the accuser should pay the legal costs of the innocent defendant. Another proposal is to subject such suits to prescreening by a government agency, as is done in Japan.

But these issues are hardly open and shut, for there is no such thing as a perfect legal system. Anything that restricts anticompetitive, private antitrust suits will almost certainly also inhibit legitimate attempts by individual firms to defend themselves from genuine acts of monopolization by rival enterprises.

SUMMARY

1. *Antitrust policy* refers to programs designed to control the growth of monopoly and to prevent big business from engaging in "undesirable" practices.
2. The *Sherman Act* is the oldest U.S. antitrust law. It prohibits contracts, combinations, and conspiracies in restraint of trade, and it also prohibits monopolization.
3. The *Clayton Act* prohibits *price discrimination* that tends to reduce competition or create monopoly; it also prohibits competing firms from sharing directors.
4. There are several other important antitrust laws, including the *Federal Trade Commission Act*, which created the commission as an independent antitrust agency, and the *Robinson-Patman Act*, which generally prohibits discriminatory price discounts.
5. The courts have usually held that control of a large share of the market by a single firm was illegal only if the firm had acquired its share by illegal means.
6. Collusion among horizontal competitors to fix prices is illegal under the Sherman Act. Firms are sometimes accused of "tacit collusion," where they cooperate without an explicit agreement, but this is difficult to prove.
7. **Predatory pricing** is pricing that is low relative to the marginal or average variable costs of the firm and that threatens to drive a competitor out of the market. Often, predatory pricing cases are initiated by competitors of the accused firm.
8. The evidence indicates that there has been no significant increase in the concentration of individual U.S. industries into fewer, relatively larger firms during the 20th century. Evidence as to whether antitrust laws have been effective in preventing monopoly is inconclusive, and observers disagree on the subject.
9. Unregulated monopoly is apt to distribute income unfairly, produce undesirably small quantities of

progressive as society deems desirable, though if marginal tax rates on rich people get extremely high, some of the potential efficiency losses might get more serious than they now seem to be. On both grounds, then, many economists—including both liberals and conservatives—view a comprehensive personal income tax as one of the best ways for a government to raise revenue.

CONCLUSION: THE VIRTUES AND VICES OF A FLAT TAX

How does the flat tax stack up against these criteria? Quite well on most, at least in its ideal form.

[6] For this reason, some economists prefer a tax on consumption to a tax on income.

- A *low* tax rate for everyone would mean that taxes would create only small tax distortions. Excess burdens would be minimal.
- A *uniform* tax rate on all types of income would eliminate tax loopholes.
- If the exemption is large enough, progressivity at the bottom of the income distribution can be maintained or even increased. For example, some plans would exempt the first $30,000 or so of income from taxation.
- As noted at the start of this chapter, filing your income-tax return under a flat tax would be extremely simple, thereby saving millions of unhappy taxpayer hours.

These are all notable virtues, which explains why the flat tax has attracted many adherents. But once we leave the realm of the *ideal* and confront the *real*, some serious problems arise.

- Some flat-tax plans would tax only *earnings*, leaving all income from capital free of tax. This would not only create strong incentives to evade the tax by transforming labor income into capital income, but it would also erode progressivity.
- It seems unlikely that Congress would eliminate *all* tax preferences. Among other things, that would mean taxing the value of employer-provided fringe benefits (like health insurance) and eliminating the deductibility of interest on home mortgages—two steps that would surely be politically unpopular.
- If large tax preferences like these remain, or if income from capital is tax free, a higher tax rate will be needed to raise the required revenue. Thus, under a realistic flat tax, most Americans might actually face a *higher* marginal tax rate than they do now.
- No flat tax can maintain progressivity at the top. Under the current tax code, the very rich pay taxes equal to about 40 percent of their taxable incomes. Under, say, a 22 percent flat tax, they would pay just 22 percent. Some Americans wonder why the very rich should get such a large tax break.

Where does this partial accounting of the pros and cons of the flat tax leave us? As usual in a serious public policy debate, with plenty of room for reasonable people to disagree! As we said back in Chapter 1, economics is not supposed to give you all the *answers*. It is supposed to teach you how to ask the right *questions*.

SUMMARY

1. Taxes in the United States have been quite constant as a percentage of gross domestic product since 1972.
2. The federal government raises most of its revenue by **direct taxes**, such as the personal and corporate *income taxes* and the payroll tax. Of these, the payroll tax is increasing most rapidly.
3. The Social Security system relied successfully on pay-as-you-go financing for decades. In recent years, however, it has been accumulating a large trust fund to be used to pay benefits to the baby boom generation when it retires. But experts do not think that trust fund will be enough.
4. State and local governments raise most of their tax revenues by **indirect taxes**. States rely mainly on sales taxes, while localities are dependent upon property taxes.
5. There is controversy over whether the property tax is **progressive** or **regressive**, and even more controversy over whether local property taxes are an equitable way to finance public education.

6. In our multilevel system of government, the federal government makes various sorts of grants to state and local governments, and states in turn make grants to municipalities and school districts. This system of intergovernmental transfers is called **fiscal federalism** and is highly controversial these days.
7. There are three concepts of fair, or "equitable," taxation that occasionally conflict. **Horizontal equity** simply calls for equals to be treated equally. **Vertical equity,** which calls for unequals to be treated unequally, has often been translated into the **ability-to-pay principle**—that people who are better able to pay taxes should be taxed more heavily. The **benefits principle** of tax equity ignores ability to pay and seeks to tax people according to the benefits they receive.
8. The **burden of a tax** is the amount of money an individual would have to be given to make her as well off with the tax as she was without it. This burden normally exceeds the taxes that are paid, and the difference between the two is called the **excess burden** of the tax.
9. Excess burden arises when a tax induces some people or firms to change their behavior. Because excess burdens signal *economic inefficiencies,* the basic principle of efficient taxation is to utilize taxes that have small excess burdens.
10. When people change their behavior on account of a tax, they often shift the burden of the tax onto someone else. This is why the **"flypaper theory of incidence"**—the belief that the burden of any tax sticks where Congress puts it—is often incorrect.
11. The burden of a sales or *excise tax* normally is shared between the suppliers and the consumers. The manner in which it is shared depends on the elasticities of supply and demand.
12. The *payroll tax* is like an excise tax on labor services. Since the supply of labor is much less elastic than the demand for labor, workers bear most of the burden of the payroll tax. This includes both the employer's and the employee's share of the tax.
13. Sometimes, "inefficient" taxes—that is, taxes that cause a good deal of excess burden—are nonetheless desirable because the changes in behavior they induce further some other social goal.
14. When there are inefficiencies in the system for reasons other than the tax system (for example, externalities), taxation can conceivably improve efficiency.

KEY TERMS

Progressive, proportional, and regressive taxes
Average and marginal tax rates
Direct and indirect taxes
Personal income tax
Payroll tax
Corporate income tax
Excise tax
Tax loopholes
Social Security system
Property tax
Fiscal federalism
Horizontal and vertical equity
Ability-to-pay principle
Benefits principle of taxation
Economic efficiency
Burden of a tax
Excess burden
Incidence of a tax
Flypaper theory of incidence
Tax shifting

QUESTIONS FOR REVIEW

1. "If the federal government continues to raise taxes as it has been doing, it will ruin the country. Americans are already overtaxed." Comment.
2. Soon after taking office in 1993, President Clinton proposed a package of tax increases, including higher income tax rates for wealthy taxpayers and higher taxes on energy. Critics argued that these taxes would harm the economy. Why did they say this?
3. Using the hypothetical income tax table just below, compute the marginal and average tax rates. Is the tax progressive, proportional, or regressive?

Income	Income Tax
$20,000	$2,000
30,000	2,700
40 000	3,200
50,000	3,500

4. Which concept of tax equity, if any, seems to be served by each of the following:
   a. The progressive income tax
   b. The flat income tax
   c. The excise tax on cigarettes
   d. The gasoline tax
5. Use the example of Mr. Figg (see the box on page 480) to explain the concepts of efficient taxes and excess burden.
6. Think of some tax that you personally pay. What steps have you taken or could you take to reduce your tax payments? Is there an excess burden on you? Why or why not?
7. Suppose the supply and demand schedules for cigarettes are as follows:

Price per Carton (dollars)	Quantity Demanded (millions of cartons per year)	Quantity Supplied (millions of cartons per year)
3.00	360	160
3.25	330	180
3.50	300	200
3.75	270	220
4.00	240	240
4.25	210	260
4.50	180	280
4.75	150	300
5.00	120	320

   a. What is the equilibrium price and equilibrium quantity?
   b. Now the government levies a $1.25 per carton excise tax on cigarettes. What is the equilibrium price paid by consumers, the price received by producers, and the quantity now?
   c. Explain why it makes no difference whether Congress levies the $1.25 tax on the consumer or the producer. (Relate your answer to the discussion of the payroll tax on pages 484–485 of the text.)
   d. Suppose the tax is levied on the producers. How much of the tax are producers able to shift onto consumers? Explain how they manage to do this.
   e. Will there be any excess burden from this tax? Why? Who bears this excess burden?
   f. By how much has cigarette consumption declined on account of the tax? Why might the government be happy about this outcome, despite the excess burden?
8. Now suppose the supply schedule is, instead:

Price per Carton (dollars)	Quantity Supplied (millions of cartons per year)
3.00	60
3.25	105
3.50	150
3.75	195
4.00	240
4.25	285
4.50	330
4.75	375
5.00	420

   a. What are the equilibrium price and equilibrium quantity in the absence of a tax?
   b. What are the equilibrium price and equilibrium quantity in the presence of a $1.25 per carton excise tax?
   c. Explain why your answer to Part b differs from Part b of the previous question, and relate this difference to the discussion of the incidence of an excise tax on pages 482–484.
9. The country of Taxmania produces only two commodities: rice and caviar. The poor spend all their income on rice, while the rich purchase both goods. Both demand for and supply of rice are quite inelastic. In the caviar market, both supply and demand are quite elastic. Which good would be heavily taxed if Taxmanians cared mostly about efficiency? What if they cared mostly about vertical equity?
10. Discuss President Reagan's statement on taxes quoted on the first page of the chapter. Do you agree with him?
11. Use the criteria of equity and efficiency in taxation to evaluate the proposal to tax capital gains at a lower rate than other sources of income.

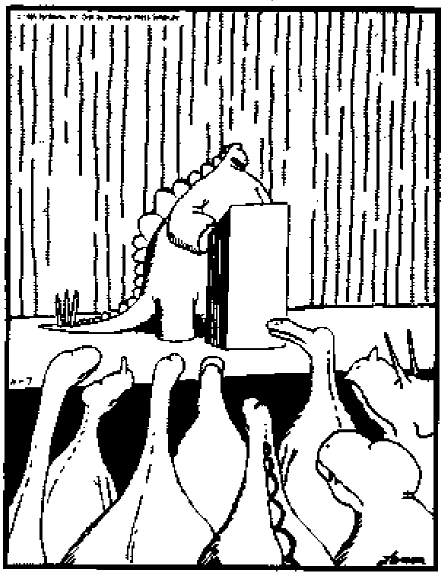
CHAPTER 21

ENVIRONMENTAL PROTECTION AND RESOURCE CONSERVATION: THE ECONOMIST'S APPROACH

Since Fuel is become so expensive, and will of course grow scarcer and dearer; any new Proposal for saving the [fuel] . . . may at least be thought worth Consideration.
Benjamin Franklin (1744)

We learned in Chapter 13 that *externalities* (the incidental benefits or damages imposed upon people not directly involved in an economic activity) can cause the market mechanism to malfunction. The first half of this chapter studies a particularly important application—externalities as an explanation of the problems of the environment. We begin by examining the environmental problems facing the world today, and ask whether everything is steadily getting worse. We then consider why an unregulated market economy generates too much pollution and analyze methods to control damage to the environment. In particular, we will look at how the price mechanism can remedy one of its principal shortcomings. In the second half of the chapter

THE FAR SIDE By GARY LARSON



"The picture's pretty bleak, gentlemen . . . The world's climates are changing, the mammals are taking over, and we all have a brain about the size of a walnut."[1]

[1]See page 494 for a brief discussion of the "greenhouse effect" and global warming.

we address a closely related subject—natural resource depletion. We will discuss the largely unfounded fears that we are running out of vital resources and learn methods for determining which key minerals truly are in short supply.

PART 1: THE ECONOMICS OF ENVIRONMENTAL PROTECTION

Environmental problems are not new. What *is* new and different is the attention the community now gives them. We can attribute much of this increased interest to rising incomes, which have reduced concerns about food, clothing, and shelter, and have allowed the luxury of concentrating on the *quality* of life.

Economic thought on the environment preceded the outburst of public concern over the subject by nearly half a century. In 1911, the British economist A. C. Pigou wrote a remarkable book called *The Economics of Welfare,* which offered an explanation of the market economy's poor environmental performance and outlined an approach to environmental policy that is still favored by most economists today and gradually seems to be winning over lawmakers and bureaucrats. Pigou's analysis suggested that a system of emissions charges can be an effective means to control pollution. In this way, the price mechanism can remedy one of its own shortcomings!

THE FACTS: IS EVERYTHING REALLY GETTING STEADILY WORSE?

First, let us see what the facts really are. Much of the discussion in the popular press gives the impression that environmental problems have been growing steadily worse, and that *all* pollution is attributable to modern industrialization and the profit system. But in fact pollution is nothing new. Medieval cities were pestholes; the streets and rivers were littered with garbage and the air stank of rotting wastes—a level of filth that was accepted as normal. And early in the 20th century, the automobile was hailed as a source of major improvement in the cleanliness of city streets, which until then had fought a losing battle against the proliferation of horse dung.

Since World War II, there has been marked progress in solving a number of pollution problems. Air quality has improved in U.S. cities and concentrations of most air pollutants are still declining. Figure 21-1 portrays the generally encouraging trends in national air pollution levels. Rapid declines in automobile pollution have played a large role in this improvement, along with decreases in emissions from power plants. There have also been some spectacular gains in water quality. In the Great Lakes region, where the Cuyahoga River once caught fire because of its toxic load and where Lake Erie was pronounced dead, tough pollution controls have gradually effected a recovery.

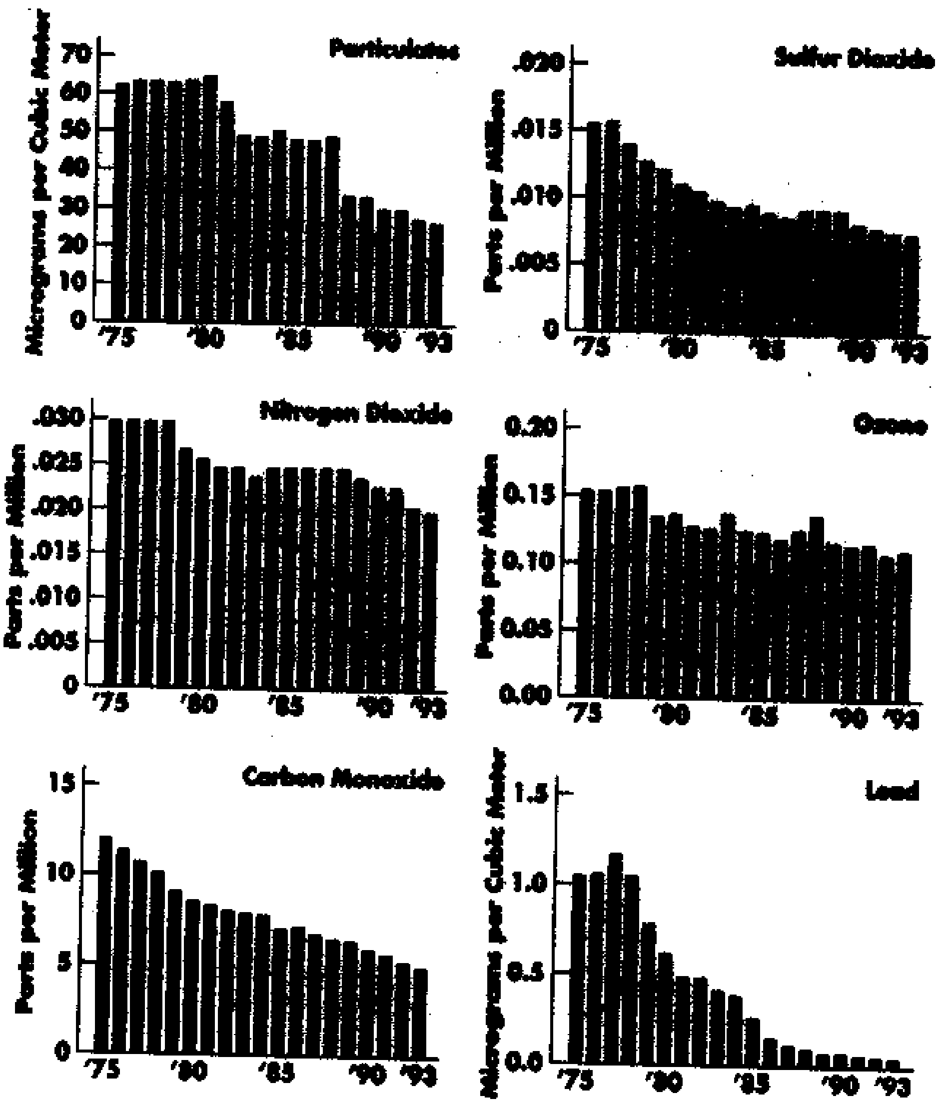
There has been progress in Europe as well. For example, the infamous, killing fogs of London, once the staple backdrop of British mystery fiction, are a thing of the past because of the improvement in air quality since 1950, and the Thames River has now been cleaned up to the point where large-scale fishing of giant conger eels has resumed after a 150-year hiatus. In short, pollution problems are not a uniquely modern phenomenon, nor is every part of the environment deteriorating relentlessly.

FIGURE 21-1 U.S. NATIONAL AIR QUALITY TRENDS, 1975–1993 (AMBIENT CONCENTRATIONS OF SIX POLLUTANTS)

Taken as a whole, the trends are encouraging. During the last two decades, the United States has made substantial progress in improving air quality. The most dramatic success story was an 89 percent reduction in lead levels in the air. Carbon monoxide has decreased by 37 percent, nitrogen oxides by 12 percent, ozone by 12 percent, and sulfur dioxide by 20 percent. Levels of particulate matter have dropped by 20 percent since 1988. With the exception of ozone, average concentrations were well below the National Ambient Air Quality Standards.

Note: After 1987, particulate matter is measured by PM10 only, an indicator of those particles smaller than 10 micrometers.

SOURCE: Council on Environmental Quality, *Environmental Quality 1993, 24th Annual Report of the CEQ* (Washington, D.C.: U.S. Government Printing Office, April 1995).



Capitalist market economies have no monopoly on pollution. Though it may seem plausible that centrally planned economies can cope better with the environmental problems resulting from externalities, the reality is that they have been environmental disasters. China, the last large communist society, has some of the world's worst air pollution, mainly from the burning of low-quality, high-sulfur coal and an almost complete lack of even the most rudimentary pollution control devices. Chinese urban smog levels are three times the average in Los Angeles, a place which Americans tend to think invented smog.[2]

Grave environmental problems also plague eastern Europe and the countries of the former Soviet Union. Poland, for example, is also one of the most polluted industrialized countries in the world, with 11 percent of its land (on which 35

[2]Worldwatch Institute, "Facing China's Limits," *State of the World 1995* (New York: W. W. Norton, 1995), pp. 123–124.

percent of its population lives) now declared "ecologically hazardous." High levels of pollution contribute to severe health problems in these countries. Life expectancies in the Czech Republic and Slovakia are the lowest in the industrialized world. The collapse of communism in the former Soviet Union revealed a staggering array of environmental horrors, including the massive poisoning of air, ground, and water in the vicinity of industrial plants. This has resulted in terrible human suffering and countless premature deaths, along with some truly monumental ecological disasters, such as the desiccation of the Aral Sea, once the fourth-largest inland sea in the world, and now reduced to less than half of its original volume.[3]

The preceding discussion is meant only to provide perspective, not to suggest that our own environment is free from problems. Despite improvements, many U.S. urban areas continue to suffer many days of unhealthful air quality, particularly during summer months. According to the U.S. Council on Environmental Quality, 59 million Americans live in counties where pollution levels in 1993 still exceeded at least one national air quality standard. (These numbers are down from 86 million in 1991.) Ozone (whose presence 12 miles up in the stratosphere protects humans from the fiercest part of the sun's ultraviolet radiation) is a serious ground-level urban pollutant. More commonly known as smog, it continues to be America's most widespread air quality problem, while carbon monoxide and particulate pollution levels also exceed federal limits in some areas. And formerly pristine wilderness areas are threatened by air pollution too.

Our world is subjected to new pollutants, some far more dangerous than those we have reduced, although less visible and less malodorous. Improperly dumped toxic substances—such as PCBs (polychlorinated biphenyls), chlorinated hydrocarbons, dioxins, heavy metals, and radioactive materials—have been found to cause cancer and threaten life and health in other ways. The danger from some of these substances can persist for thousands of years, causing damage that is all but irreversible.

Even these problems pale when compared to an uncertain, but very real, environmental threat that hangs over our future—the long-term warming of the earth's atmosphere. A growing body of evidence suggests that the documented warming of the last century, and especially the last few years, is at least partly the result of human activities that have increased the "greenhouse gases" in the atmosphere. The buildup of carbon dioxide from the burning of fossil fuels such as oil, natural gas, and coal is a particular problem. Figure 21-2 charts the warming trend. Forecasts of future warming range from 1.8° to 6.3° Fahrenheit by the year 2100, a dramatic change that may shift world rain patterns, disrupt agriculture, threaten coastal cities with inundation, and expand deserts.[4]

While environmental problems are neither new nor confined to capitalist, industrialized economies, we continue to inflict damage on ourselves and our surroundings.

---

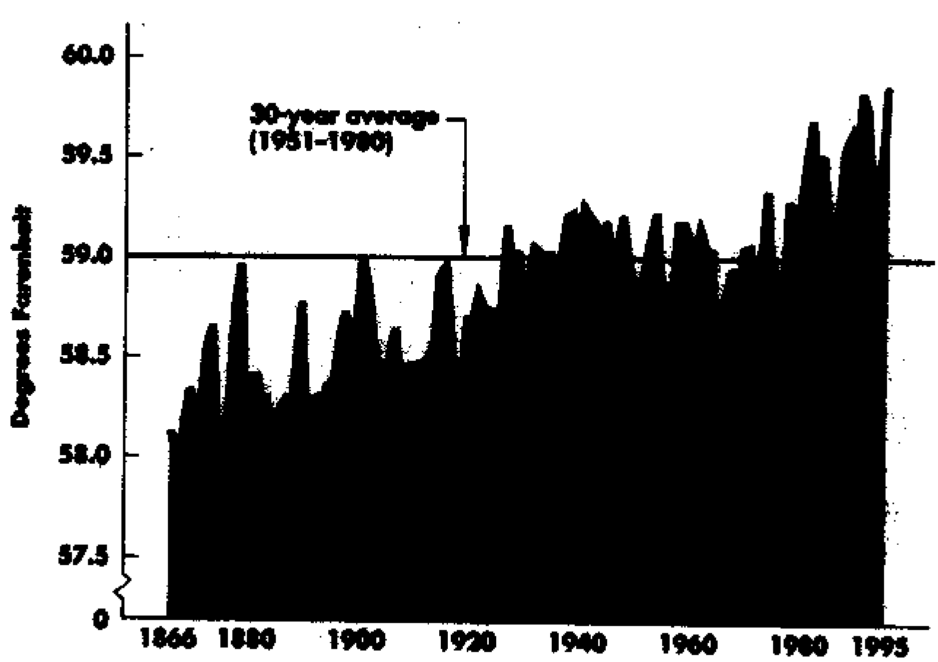
[3]World Resources Institute, *The 1994 Information Please Environmental Almanac* (Boston: Houghton Mifflin, 1994); Marlise Simons, "East Europe Sniffs Freedom's Air, and Gasps," *The New York Times*, November 3, 1994, pp. A1 and A14; Murray Feshbach and Alfred Friendly, Jr., *Ecocide in the USSR: Health and Nature under Siege* (New York: Basic Books, 1992).

[4]Thomas R. Karl et al., "Trends in U.S. Climate during the Twentieth Century," *Consequences: The Nature and Implications of Environmental Change* 1, no. 1 (Spring 1995) (U.S. National Oceanic and Atmospheric Administration); "Global Warming Experts Call Human Role Likely," *The New York Times*, September 10, 1995, pp. A1 and A8.

FIGURE 21-2 GLOBAL SURFACE AIR TEMPERATURE, ANNUAL AVERAGE, 1866–1995

The atmospheric warming of the last century may turn out to be at least partly attributable to human activities such as the burning of fossil fuels.

SOURCE: "Global Warming Experts Call Human Role Likely," *The New York Times*, September 10, 1995, pp. A1 and A8, which cites National Aeronautics and Space Administration, Goddard Institute for Space Studies.



THE LAW OF CONSERVATION OF MATTER AND ENERGY

The physical law of conservation of matter and energy tells us that objects cannot disappear—at most they can be changed into something else. Oil, for instance, can be transformed into heat (and smoke) or into plastic—but it will never vanish. This means that after a raw material has been used, either it must be used again (recycled) or it becomes a waste product that requires disposal.

If it is not recycled, any input used in production must ultimately become a waste product. It may end up in some municipal dump, or it may literally go up in smoke, contributing to atmospheric pollution, or it may be transformed into heat, warming up adjacent waterways and killing aquatic life. But the laws of physics tell us nothing can be done to make used inputs disappear altogether.

We create an extraordinary amount of solid waste—each American produced an average of 4.3 pounds of trash every day in 1990 (up from 4.2 pounds in 1988), despite our efforts to reduce the amount of this waste. Fortunately, in the face of this rising tide of garbage, recycling rates for many commonly used materials (like aluminum, paper, and glass) are rising in the United States and many other industrial countries. The Organization for Economic Cooperation and Development reports 1990 recycling rates of 12.5 percent for Sweden, 10 percent for Canada, 6 percent for the United Kingdom, 4 percent for France, and 3 percent for Japan. Here in the United States the Environmental Protection Agency estimates that the rate at which materials in the municipal solid waste stream were recycled rose from 7 percent in 1960 (6 million tons) to 17 percent (33 million tons) in 1990. Particularly in the last few years, there has been a remarkable increase in community recycling programs, with a fivefold increase in local programs since 1988.

In the authors' own town of Princeton, New Jersey, recycling rates now top 50 percent of municipal solid waste, a rate that is not ur [illegible]

and state ordinances have sometimes helped these efforts along by charging residents for each bag of trash sent to a landfill. Perkasie, Pennsylvania, reduced the waste load by 41 percent and Ilion, New York, by 37 percent after such systems were initiated.

Yet there have been problems; for example, the market for recyclable materials has fluctuated dramatically in the last decade, with the bottom dropping out of the newsprint market in the early 1990s as local recycling programs flooded the market with vast quantities of paper. At one point the problem was so bad that some U.S. cities had to pay recyclers to take old newspapers off their hands. But the trend began reversing after the opening of 85 new U.S. paper mills that use recycling technology. As a result, the cost of old newspapers skyrocketed 1,338 percent between 1993 and 1995. During the same period, the price paid by processors for used, clear glass containers jumped 78 percent, as did the value of discarded aluminum cans.[5]

GOVERNMENTS AND INDIVIDUALS AS DAMAGERS OF THE ENVIRONMENT

Many people think of industry as the primary villain in environmental damage. But:

While business firms have done their share in harming the environment, private individuals and government have also been major contributors.

Individual car owners are responsible for much of the air pollution in cities; wood-burning stoves are a source of particulate pollution, and wastes from flush toilets and residential washing machines also cause significant harm.

Governments, too, add to the problem. The wastes of municipal treatment plants are a major source of water pollution. Military aircraft create exhaust and cause noise pollution. Obsolete atomic materials and by-products associated with chemical and nuclear weapons are among the most dangerous of all wastes, and their disposal remains an unsolved problem.

Governments also construct giant dams and reservoirs that flood farmlands and destroy canyons, often rendering surrounding soil unusable through seepage of salt into the earth. Drainage of swamps has altered local ecology irrevocably; canal-building has diverted the flow of rivers. The U.S. Army Corps of Engineers has been accused of acting on the basis of this so-called *edifice complex*. But that complex reached its greatest heights in the communist states under Stalin, whose pride in enormous (but environmentally destructive) hydroelectric installations and huge canals was well publicized in the Soviet press.

ENVIRONMENTAL DAMAGE AS AN EXTERNALITY

Our very existence makes some environmental damage inevitable. In order to eat and protect themselves from the elements, people must use up the earth's resources and generate wastes.

---

[5]Organisation for Economic Cooperation and Development, *OECD Environmental Data 1993* (Paris: OECD, 1993), p. 141; "Increase in Recycling for Mercer County," *Town Topics* (Princeton, New Jersey), July 12, 1995, p. 9; "The Best," *Time*, January 2, 1995; World Resources Institute, *1994 Environmental Almanac*; "Turning Trash into Cash," *U.S. News & World Report*, July 17, 1995, p. 43.

Environmental damage cannot be reduced to zero. As long as the human race survives, it is literally impossible to eliminate such damage completely.

The real issue then is not whether pollution should exist at all, but whether environmental damage in an unregulated market economy tends to be more serious and widespread than the public interest can tolerate. This issue immediately raises three key questions. First, why do economists believe that environmental damage is unacceptably severe *in terms of the public interest?* Second, why does the market mechanism, which is so good at providing about the right number of running shoes and refrigerators, generate too much pollution? And, third, what can we do about it? We will consider these questions in order.

Economists do not claim any special ability to judge what is good for the public. They usually accept people's wishes as "the public interest." When the economy responds to these wishes as closely as the available resources and technology permit, economists conclude it is working effectively. When it operates in a way that frustrates the desires of the people, they conclude that the economy is functioning improperly. In such terms, why do economists believe the market generates "too much" pollution?

In Chapter 13 we discussed externalities as a primary source of failure of the market mechanism. Recall that *externalities* are the incidental benefits or damages imposed upon people not directly involved in an economic activity. The emission of pollutants constitutes one of the most clear-cut examples. The toxic fumes from a chemical plant affect not only the employees of the plant and its customers, but other people as well. Because the firm does not pay a price for this incidental damage, the owners of the firm have no financial incentive to limit emissions, particularly since emission control costs money. Instead, they will find it profitable to continue their toxic emissions as though the fumes caused no external damage to the community.

This is a *failure of the pricing system*—the pollution-generating business firm is able to use up some of the community's clean air without paying for the privilege. Just as the firm would undoubtedly use oil and electricity wastefully if they were available at no charge, the firm will use "free" air wastefully, despoiling it with chemical fumes far beyond the level that the public interest can justify.

EXTERNALITIES

Externalities play a crucial role affecting the quality of life. They show why the market mechanism, which is so efficient in supplying consumers' goods, has a much poorer record in terms of its effects on the environment. The problem of pollution illustrates the importance of externalities for public policy and indicates why their analysis is one of our **Ideas for Beyond the Final Exam.**

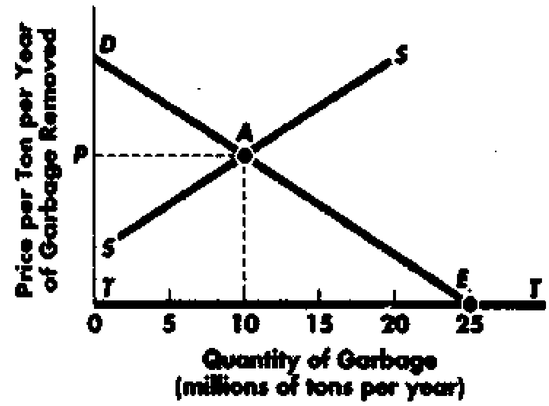
SUPPLY-DEMAND ANALYSIS OF ENVIRONMENTAL EXTERNALITIES

We can use basic supply-demand analysis to explain both how externalities lead to environmental problems and how these problems can be cured. As an illustration, consider the damage that massive generation of garbage does to our environment.

Figure 21-3 shows a demand curve, *DE*, for garbage removal. As usual, this curve has a negative slope, meaning that if the price of garbage removal rises, people will demand less garbage removal. They may take more waste to recycling centers, repair broken items rather than throwing them out, and so on.

FIGURE 21-3 FREE DUMPING OF POLLUTANTS AS AN INDUCEMENT TO ENVIRONMENTAL DAMAGE

Wastes impose costs upon the community. If the polluter pays nothing for this damage, the wastes are, in effect, removed with zero charges to the polluter (blue removal supply curve *TT*). The polluter is induced to create large amounts of pollution (25 million tons). If the charge to the polluter reflects the true cost to the community (supply curve *SS*), it will pay the polluter to emit only 10 million tons.



The graph also shows the supply curve, *SS*, of an ideal market for garbage removal. As we saw in our analysis of competitive industries (Chapter 9), the position of the market's long-run supply curve is the average cost of garbage removal. If suppliers had to pay the full costs of garbage removal—including the cost of the pollution caused when the garbage is burned at the dump—the supply curve would be comparably high (as drawn in the graph). For the community depicted in the graph, the price of garbage removal is *P* dollars per ton, and 10 million tons is generated (point *A*).

But what if the community's government decides to remove garbage "for free"? Of course, the consumer still really pays through taxes, but not in a way that makes each household pay for the quantity of garbage that it actually produces. The result is that the supply curve is no longer *SS*. Rather it becomes the blue line *TT*, which lies along the horizontal axis, because any household can increase the garbage it throws away at no cost to itself. Now the intersection of the supply and demand curve is no longer point *A*. Rather it is point *E*, at which the price is zero, and the quantity of garbage generated is 25 million tons—a substantially greater amount.

Similar problems occur if the community offers the dissolved oxygen in its waterways and the purity of its atmosphere without charge. The amount that will be wasted and otherwise used up is likely to be enormously greater than it would be if users had to pay for the cost of their actions to society. That is a key reason for the severity of our environmental problems.

The magnitude of our pollution problems is attributable in large part to the fact that the market lets individuals, firms, and government agencies deplete such resources as dissolved oxygen in the water and pure air without charging them any money for using those resources.

It follows that one way of dealing with pollution problems is to charge those who emit pollution, and despoil the environment in other ways, a price commensurate with the costs they impose on society.

BASIC APPROACHES TO ENVIRONMENTAL POLICY

In broad terms, three general methods have been proposed for the control of activities that damage the environment.

1. *Voluntary programs,* such as nonmandatory investment in pollution-control equipment by firms motivated by social responsibility, or voluntary recycling of solid wastes by consumers.
2. *Direct controls,* which either (a) impose legal ceilings on the amount any polluter is permitted to emit or (b) specify how particular activities must be carried on—for example, they may prohibit backyard incinerators or the use of high-sulfur coal or require smokestack "scrubbers" to capture emissions of power plants.
3. *Taxes on emissions,* or the use of other monetary incentives or penalties to make it unattractive financially for emitters of pollutants to continue to pollute as usual.

All of these methods have useful roles. Let us consider each of them in turn.

VOLUNTARISM

Voluntarism often has proved weak and unreliable. Strictly voluntary programs for recycling of garbage have rarely managed to keep more than a very small fraction of a community's wastes from the garbage dump. Some well-intentioned business firms have made sincere attempts to adopt environmentally beneficial practices. Yet competition has usually prevented them from spending more than token amounts for this purpose. No business, whatever its virtues, can long afford to spend so much on good works that rivals can easily underprice it. As a result, voluntary business programs sometimes have been more helpful to the companies' public relations activities than to the environment. Firms with a real interest in environmental protection have called for legislation that *requires* all firms, including competitors, to undertake the same measures, thereby subjecting all firms in the industry to similar cost handicaps.

Yet voluntary measures do have their place. They are appropriate where surveillance and, consequently, enforcement is impractical, as in the prevention of littering by campers in isolated areas, where there is no alternative to an appeal to people's consciences. And in brief but serious emergencies, in which there is no time to plan and enact a systematic program, there also may be no good substitute for voluntary compliance.

Several major cities have, for example, experienced episodes of temporary but dangerous concentrations of pollutants, forcing the authorities to appeal to the public to cut emissions drastically. The public response to appeals requiring cooperation for short periods often has been enthusiastic and gratifying, particularly when civic pride was a factor; for example, during the 1984 Summer Olympic Games, Los Angeles city officials asked motorists to car pool, businesses to stagger work hours, and truckers to restrict themselves to essential deliveries and to avoid rush hours. The result was an extraordinary decrease in traffic and smog, such that the 6,000-foot San Gabriel Mountains suddenly became visible behind the city. To summarize:

Voluntary programs are not dependable ways to protect the environment. However, in brief, unexpected emergencies or where effective surveillance is impossible, the policymaker may have no other choice.

DIRECT CONTROLS

Direct controls have been the chief instrument of environmental policy in the United States (the so-called "command and control" approach). The federal government, through the Environmental Protection Agency (EPA), formulates

standards for air and water quality and requires state and local governments to adopt rules that will ensure achievement of those goals. Probably the best known of these are the standards for automobile emissions. New automobiles are required to pass tests showing that their emissions do not exceed specified amounts.

As another example, localities sometimes prohibit the use of particularly "dirty" fuels by industry or require the adoption of processes to "cleanse" those fuels.

TAXES ON EMISSIONS

Most economists agree that a nearly exclusive reliance on direct controls is a mistake and that, in most cases, financial penalties on polluters can do the same job more dependably, effectively, and économically.

The most common suggestion is that firms be permitted to pollute all they want but be forced to pay a tax for the privilege, in order to make them *want* to pollute less. Under such a plan, the quantity of the polluter's emissions would be metered just like the use of electricity. At the end of the month the government would automatically send the polluter a bill charging a stipulated amount for each gallon (or other unit) of emissions. (The amount would also vary with the quality of the emissions—a higher tax rate being imposed on emissions that were more dangerous or unpleasant.) Thus, the more environmental damage done, the more the polluter would have to pay. Such taxes are deliberately designed to *encourage* the use of a glaring tax loophole—the polluter *can* reduce the tax owed by polluting less. In terms of Figure 21-3, if the tax is used to increase the payment for waste emissions from zero (blue supply line *TT*) and instead forces the polluter to pay its true cost to society (line *SS*), emissions will automatically be reduced from 25 million to 10 million tons.

Businesses do respond to such taxes. One widely publicized example is the Ruhr River basin in Germany, where emissions taxes have been used for decades. Though the Ruhr is one of the world's most concentrated industrial centers, the rivers that are protected by taxes are clean enough for fishing and other recreational purposes. Firms have also found it profitable to avoid taxes by extracting pollutants from their liquid discharges and recycling them.

EMISSIONS TAXES VERSUS DIRECT CONTROLS

It is important to see why taxes on emissions may prove more effective and reliable than direct controls. Direct controls essentially rely on the enforcement mechanism of the criminal justice system. But the polluter who violates the rules must first be caught. Then the regulatory agency must decide whether it has enough evidence to prosecute. Next, it must win its case in court. And, finally, the court must impose a penalty strong enough to matter. If any *one* of these does not occur, the polluter gets away with the environmentally damaging activities.

ENFORCEMENT ISSUES

The enforcement of direct controls requires vigilance and enthusiasm by the regulatory agency, which must assign the resources and persons needed to carry out the task of enforcement. In many cases the resources devoted to enforcement are pitifully small, or at least tend to wax and wane with the political fortunes

of the regulatory agency. For instance, during President Reagan's administration, environmental outlays were cut severely and the progress that had been made in the 1970s nearly came to a standstill. In more recent times, President Bill Clinton's plans to devote considerably greater resources and human power to environmental matters have been jeopardized by a Congress intent on decreasing EPA funding drastically.

The effectiveness of direct controls also depends upon the speed and rigor of the courts. Yet the courts are often slow and lenient. The notorious case of the Reserve Mining Company is one example. It took more than a decade of litigation to stop this company from pouring its wastes (which contain asbestos-like fibers believed to cause cancer) into Lake Superior, the source of drinking water for a number of communities.

Finally, direct controls can work only if the legal system imposes significant penalties on violators. Lately, there have been some cases of significant penalties: In what may be the most dramatic instance, Exxon has paid $1.1 billion in fines for the oil spill damage caused to Alaska's Prince William Sound when the tanker *Exxon Valdez* ran aground, and faces the possibility of much higher penalties for this incident. And some polluters have even served prison terms for their environmental misdeeds. But there are many more cases in which large firms have been convicted of polluting and fined amounts that are beneath the notice of even a relatively small corporation.[6]

In contrast to all this, pollution taxes are automatic and certain. No one need be caught, prosecuted, convicted, and punished. The tax bills are just sent out automatically by the untiring tax collector. The only sure way for the polluter to avoid paying pollution charges is to pollute less.

EFFICIENCY IN CLEANUP

A second important advantage of emissions taxes is that they do the job at a lower cost than direct controls. Statistical estimates for several pollution-control programs suggest that the cost of doing the job through direct controls can easily be twice as high as under the tax alternative. Why should there be such a difference? The answer is that under direct controls, emissions cutbacks are usually *not* apportioned among the various firms on the basis of ability to reduce pollution cheaply and efficiently.

Suppose it costs Firm A only 3 cents a gallon to reduce emissions while Firm B must spend 20 cents a gallon to do the same job. If each firm spews out 2,000 gallons of pollution a day, authorities can achieve a 50 percent reduction in pollution by ordering both firms to limit emissions to 1,000 gallons a day. This may or may not be fair, but it is certainly not efficient. The social cost will be 1,000 times 3 cents, or $30, to Firm A and 1,000 times 20 cents, or $200, to Firm B, a total of $230. If, instead, the government had imposed a tax of 10 cents a gallon, Firm A would have done all the work at a lower cost. Let us see why. Firm A would have cut its emissions out altogether, paying the 3 cents a gallon this required to avoid the 10-cents-a-gallon tax. Firm B would have gone on polluting as before, because it would be cheaper to pay the tax than the 20 cents a gallon it would cost to control its pollution. In this way, under the tax, *total daily emissions would still be cut by 2,000 gallons a day.* But the total daily cost of the

[6] "Corporate Liability; UnExxonerated," *The Economist*, June 18, 1994, p. 32; "Oil Polluters Get Prison Terms," *ClariNet e.News* (http://www.clari.net), October 12, 1994.

program would therefore be $60 (3 cents × 2,000 gallons) instead of the $230 it would cost under direct controls.

The secret of the efficiency induced by a tax on pollution is straightforward. Only polluters who can reduce emissions cheaply and efficiently can afford to take advantage of the built-in loophole—the opportunity to save on taxes by reducing emissions. The tax approach simply assigns the job to those who can do it most effectively—and rewards them by letting them escape the tax.

ADVANTAGES AND DISADVANTAGES

Given all these advantages of the tax approach, why would anyone want to use direct controls?

There are three general and important situations in which direct controls have a clear advantage:

1. *Where an emission is so dangerous that it is prohibited altogether.*
2. *Where a sudden change in circumstances—for example, a dangerous air-quality crisis—calls for prompt and substantial changes in conduct, such as temporary reductions in use of cars.* It is difficult and clumsy to change tax rules, and direct controls will usually do a better job here. The mayor of a city threatened by a dangerous air-quality crisis can, for example, forbid the use of private passenger cars until the crisis passes.
3. *Where effective and dependable metering devices have not been invented or are prohibitively costly to install and operate.* In such cases there is no way to operate an effective tax program because authorities cannot determine the level of emissions and so cannot calculate the tax bill. In that case the only effective option may be to *require* firms to use "clean" fuel, or install emissions-purification equipment.

OTHER FINANCIAL DEVICES TO PROTECT THE ENVIRONMENT: EMISSIONS PERMITS

The basic idea underlying the emissions-tax approach to environmental protection is that it provides financial incentives that induce polluters to reduce the damage they do to the environment. But there is at least one other form of financial inducement that deserves consideration: requiring polluters to buy *emissions permits* that authorize the emission of a specified quantity of pollutant. Such permits would be offered for sale in limited quantities fixed by the authorities at prices set by demand and supply.

Under this arrangement, the environmental agency decides what quantity of emissions per unit of time (say, per month) is tolerable and then issues a batch of permits authorizing (altogether) just that amount of pollution. The permits are sold to the highest bidders. The price is therefore determined by demand and supply. It will be high if the number of permits offered for sale is small and a large number of firms need permits in order to carry out their industrial activities. Similarly, the price of a permit will be low if authorities issue many permits but the quantity of pollution for which they are demanded is small.

The emissions permit in many ways works like a tax—it simply makes it too expensive for firms to continue polluting as much as before. However, the permit approach has some advantages over taxes. For example, it reduces uncer-

tainty about the quantity of pollution that will be emitted. Under a tax, we cannot be sure about this in advance, since that depends on the extent to which polluters respond to a given tax rate. In the case of permits, environmental authorities simply decide on an emissions ceiling in advance and then issue permits authorizing just that quantity of emissions.

Many people react indignantly to the notion of emissions permits, calling them "licenses to pollute." Yet the EPA has introduced some compromise measures that seem palatable politically and that can be regarded as approximations to a market in emissions permits. (The box below describes the recently inaugurated market in sulfur dioxide emissions permits that promises to cut both emissions and polluter costs.)

A MARKET APPROACH TO CUTTING ENVIRONMENTAL PROTECTION COSTS

To reduce the damage that acid rain causes to lakes, streams, buildings, and visibility throughout the eastern United States and Canada, the 1990 Clean Air Act established the largest market ever devised for trading in emissions permits. The law mandates a 50 percent cut in annual sulfur dioxide (SO_2) emissions from electric utilities—from 18 million tons per year to 9 million—and lets market forces guide utilities to the least-cost method for achieving the overall cut.

Here's how it works. The program, administered by the EPA, distributes a limited number of tradeable emission permits, called "allowances," to electric power plants and requires plants to hold enough allowances each year to cover their emissions. The EPA also maintains trading records, verifies that plants hold enough allowances, and levies a stiff per-ton fine if a plant exceeds its pollution limit.

Instead of giving clean air away free, the acid rain program puts a known cap on pollution and gives property-like ownership to a resource that previously belonged to everybody—but nobody. Unlike traditional regulations, this program achieves a specified emissions cut and the market rewards those who find ways to cut pollution more efficiently. Society gets cleaner air at lower cost, meaning smaller electric-rate increases for consumers and preserved competitiveness (and jobs) for American businesses. Environmentalists win too; since the program makes it cheaper to cut emissions, Congress was more willing to impose larger overall reductions.

Unlike direct controls, the market provides a profit incentive to cut emissions most cheaply, rewards innovation, introduces competition among the various pollution-control methods, and relies on the expertise of plant managers, not government officials, to figure out the best way to cut pollution. Those who make extraordinary emission cuts can sell their unneeded allowances. For example, if a plant cuts emissions for \$70 per ton and sells unneeded allowances for \$130 per ton, it will realize a \$60 profit. If the plant finds a way to cut emissions at even lower cost, its profits will rise.

The flexibility lets plant operators choose from numerous pollution-control options based on cost: installing emission scrubbers; switching to cleaner fuels, like low-sulfur coal or natural gas; inducing energy conservation; or hiring another plant to make extra emission cuts on their behalf by purchasing allowances.

To stimulate the market, the program requires yearly environmentally neutral allowance auctions. The auctions, administered by the Chicago Board of Trade, generate the public price signals that utilities need to identify the cheapest compliance option. Anyone can buy the allowances, and the auctions make buying simple for those who wish to "retire" allowances in order to force greater overall emission cuts. Groups ranging from school children to environmental organizations have put their money where their mouth is by exploiting this unprecedented option for directly improving the environment.

At the 1996 auction, the price of a one-ton SO_2 permit cleared at just \$66, down from the previous year's price of \$132 (and far below the \$300–\$500 per-ton compliance cost predicted for a traditional, "command and control" approach). The fall in the price of these permits reflected the low demand for them, as utilities increasingly found it profitable to use less-polluting low-sulfur coal, thus requiring fewer SO_2 permits to operate. This has all added up to savings in the billions of dollars in pollution control costs.

SOURCE: Contributed by Michael Walsh, senior vice president, Centre Financial Products, Chicago; formerly senior economist at the Chicago Board of Trade.



TWO CHEERS FOR THE MARKET

We have seen in the first part of this chapter that protecting the environment is one task that cannot be left to the free market. Because of the important externalities involved, the market will systematically allocate too few resources to the job. However, this market failure does not imply that we must disregard the price mechanism. On the contrary, we have seen that a legislated market solution based on pollution charges may well be the best way to protect the environment. At least in this case, the power of the market mechanism can be harnessed to correct its own failings.

We turn now, in the second half of the chapter, to the use of natural resources, where the market mechanism also plays a crucial role.

PART 2: THE ECONOMICS OF ENERGY AND NATURAL RESOURCES

The "energy crisis" of the 1970s and early 1980s, when oil prices jumped dramatically, had profound effects throughout the world—one of which was the end of the belief that the stock of natural resources was all but unlimited and simply ours for the taking. Indeed, at that time there was near-panic about the prospect of exhaustion of a number of resources. The front page of a leading magazine asked, "Are we running out of *everything?*"

Natural resources have always been scarce, and they undoubtedly have been used wastefully. Nevertheless, we are *not* about to run out of the most vital resources. There is reason to be optimistic about the availability of substitutes, and many of the shortages of the 1970s can justifiably be ascribed as much to the folly of government programs as to imminent exhaustion of petroleum and other natural resources.

A PUZZLE: THOSE RESILIENT RESOURCE SUPPLIES

It is a plain fact that the earth is endowed with only finite quantities of such vital resources as oil, copper, lead, and coal. This has elicited many worried forecasts about the inevitable, and imminent, exhaustion of one resource or another. The box on the next page lists a number of bleak prophecies about oil production in the United States, all of which proved far off the mark.

And, in fact, far from running out, supplies of many key minerals and fuels are growing. In recent decades known supplies of most minerals have generally grown at least as fast as production, and in many cases have far outstripped production. For example, in 1990 world reserves of tin were estimated at 5.9 million metric tons (mmt). During 1991, 203,000 metric tons of tin were mined from the earth. Nonetheless, by 1992 world reserves of tin had *risen* to 8 mmt. For bauxite (the ore used to produce aluminum), 1990 reserves were 21,600 mmt, 1991 production topped 100 mmt, and 1992 reserves rose to 23,000 mmt. A similar odd story can be told for crude oil, coal, zinc, and many more.[7] Economic principles, as we will see at the end of this chapter, help a great deal in clearing up these mysteries.

---

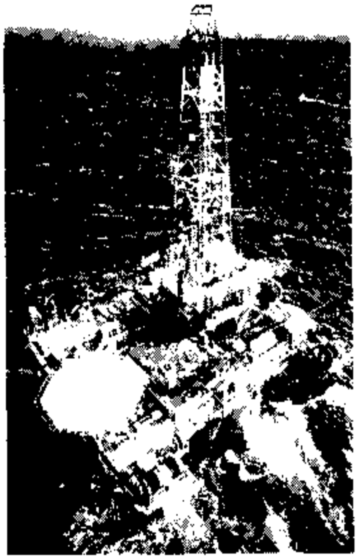
[7]World Resources Institute, *World Resources 1994–95* (New York: Oxford University Press, 1994) and other issues.

THE PERMANENT FUEL CRISIS

Humanity has a long history of panicking about the imminent exhaustion of natural resources. In the 13th century a large part of Europe's forests was cut down, primarily for use in metalworking (much of it for armor). Wood prices rose, and there was a good deal of talk about depletion of fuel stocks. People have been doing it ever since, as the following cases illustrate.

Past Petroleum Prophecies (and Realities)

Date	U.S. Oil Production Rate (billion barrels/year)	Prophecy	Reality
1866	0.005	Synthetics are available if oil production should end (U.S. Revenue Commission).	In the next 82 years, the United States produced 37 billion barrels with no need for synthetics.
1891	0.05	Little or no chance for oil in Kansas or Texas (U.S. Geological Survey).	14 billion barrels produced in these two states since 1891.
1914	0.27	Total future production only 5.7 billion barrels (Official of U.S. Bureau of Mines).	34 billion barrels produced since 1914 or six times this prediction.
1920	0.45	United States needs foreign oil and synthetics: peak domestic production almost	1948 U.S. production in excess of U.S. consumption and more than four times



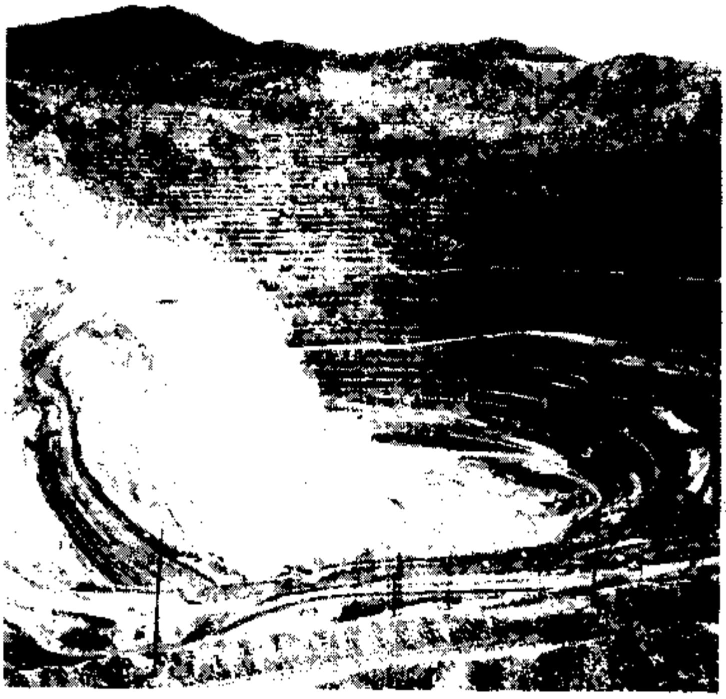
SOURCE: William M. Brown, "The Outlook for Future Petroleum Supplies," in Julian L. Simon and Herman Kahn.

"GREEN" GDP

GDP as conventionally measured does not net out the costs of pollution or the depletion of renewable resources like forests, nor does it "charge" for using up nonrenewable mineral resources. The consequence, some critics argue, is that GDP overstates economic well-being. The government is currently developing experimental measures that will adjust for each of these factors, and the first phase of its research turned up some surprising results.

Statisticians valued the U.S. stock of mineral resources annually from 1958 to 1991. The idea was to estimate how much had been consumed. However, despite the fact that the U.S. economy used enormous amounts of minerals over the 33-year period, the statisticians discovered that proven supplies barely declined. How can this be?

In a purely physical sense, of course, it cannot be. America certainly had fewer actual mineral deposits in the ground in 1991 than in 1958. However, new sources of minerals were discovered during the period, and new technologies made it possible to mine in places previously viewed as either inaccessible or uneconomic. When mineral use was netted against new discoveries, *proven* reserves barely fell over the 33-year period. Therefore, making an adjustment for the depletion of mineral resources would hardly change the official GDP numbers at all.

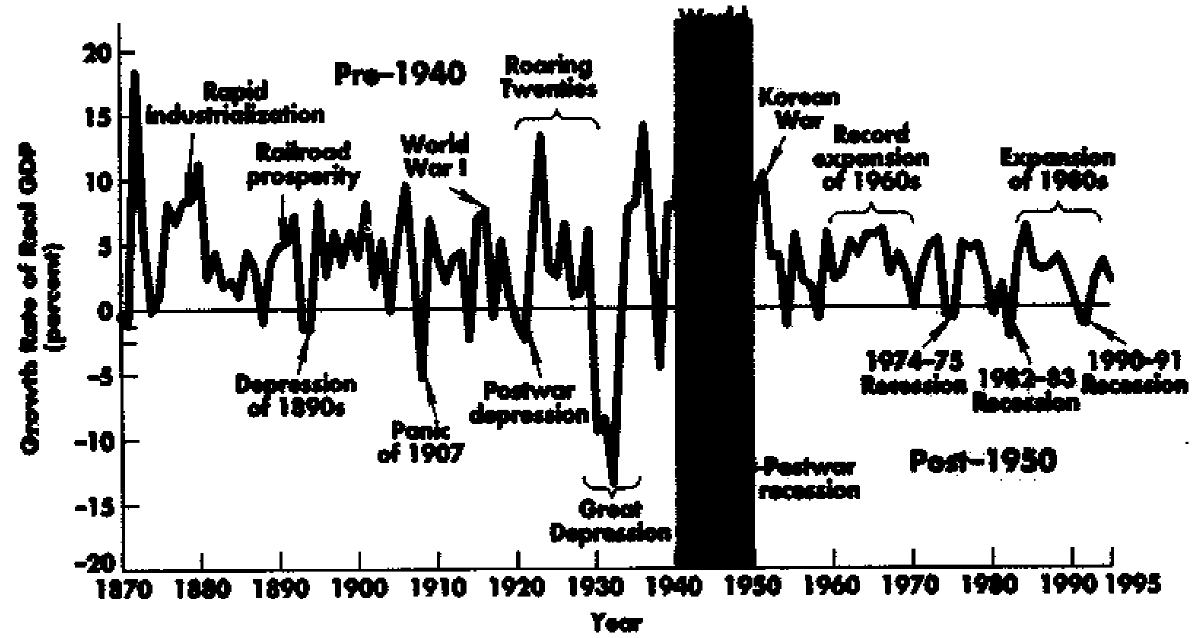


In sum, although both real GDP and the price level have grown a great deal over the past 125 years, neither has grown smoothly. The ups and downs of both real growth and inflation have been important economic events that need to be explained. Parts VI and VII develop a macroeconomic theory designed to do precisely that.

As you look at these graphs, the Great Depression of the 1930s is bound to catch your eye. The decline in economic activity from 1929 to 1933 indicated in Figure 22-3 was the most severe in our nation's history, and the rapid deflation in Figure 22-4 was most unusual. The Depression is but a dim memory now, but those who lived through it—including some of your grandparents—will never forget it.

Statistics usually conceal the true drama of economic events. But this is not so of the Great Depression—instead they stand as bitter testimony to its severity. The production of goods and services dropped 30 percent, business investment almost dried up entirely, and the unemployment rate rose ominously from about 3 percent in 1929 to 25 percent in 1933. One person in four was jobless. From the data alone, you can conjure up pictures of soup lines, beggars on street corners, closed factories, and homeless families. (See the box on page 529.)

FIGURE 22-3 THE GROWTH RATE OF REAL GROSS DOMESTIC PRODUCT OF THE UNITED STATES, 1870–1995



This time-series chart displays the growth rate of real gross domestic product in the United States from 1870 to 1995. (Here real GDP is measured in 1992 prices.) The Great Depression (1929–1939) stands out vividly. The years during and just after World War II are shaded. Does the growth rate look smoother to the right of this shaded area?

SOURCE: Constructed by the authors from Commerce Department data for 1929–1995. Data for 1869–1928 are based on research by Professor Christina Romer.

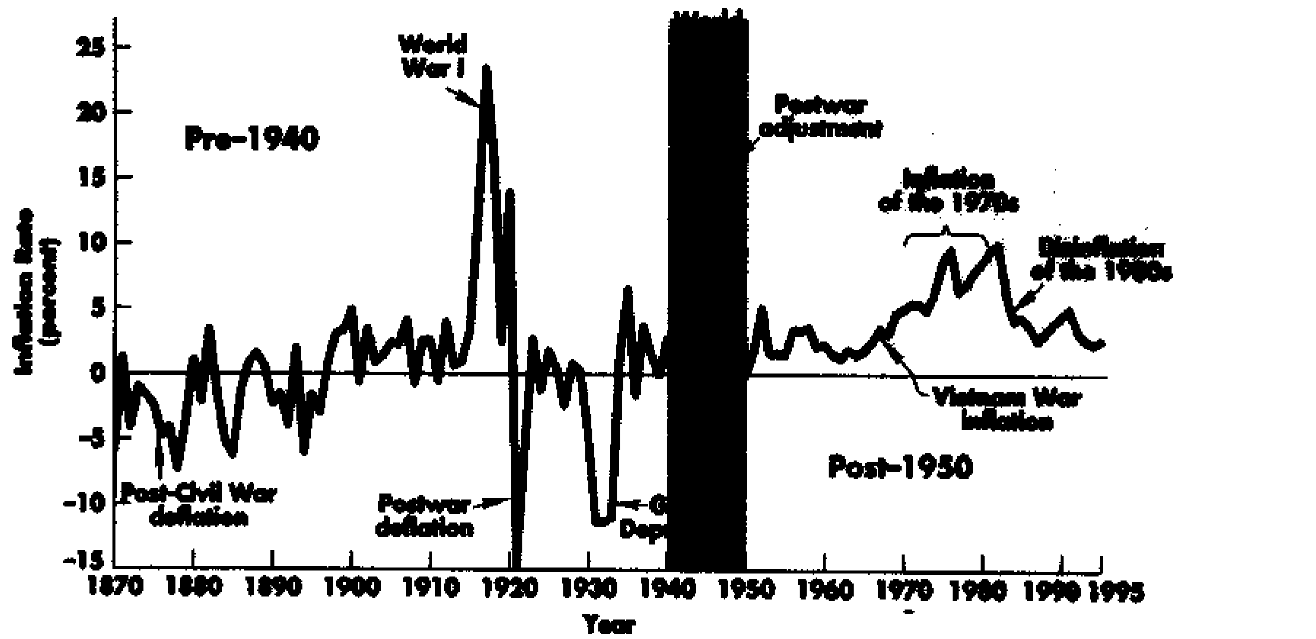
The Great Depression was a worldwide event. No country was spared its ravages, which literally changed the histories of many nations. In Germany, it facilitated the ascendancy of Nazism. In the United States, it enabled Franklin Roosevelt's Democratic party to engineer one of the most dramatic political realignments in history and to push through a host of political and economic reforms.

The worldwide depression also caused a much-needed revolution in the thinking of economists. Up until the 1930s, the prevailing economic theory held that a capitalist economy occasionally misbehaved but had a natural tendency to cure recessions or inflations by itself. The roller coaster bounced around but did not normally run off the tracks. This optimistic view was not confined to academia. Most political and business leaders shared it, as well. As the great American humorist Will Rogers remarked with characteristic sarcasm:

> It's almost been worth this depression to find out how little our big men knew. Mayby [sic] this depression is just "normalcy" and we don't know it. It's made a dumb guy as smart as a smart one.... Depression used to be a state of mind, now it's a state of coma, now it's permanent. Last year we said, "Things can't go on like this," and they didn't, they got worse.[6]

[6]From *Sanity Is Where You Find It* by Will Rogers, edited by Donald Day; copyright 1955 by Rogers Company; reprinted by permission of Houghton Mifflin Company; pages 120–121.

FIGURE 22-4 THE INFLATION RATE IN THE UNITED STATES, 1870–1995



This time-series chart portrays the behavior of the U.S. inflation rate from 1870 to 1995. (The specific price index used is called the *GDP deflator*, and it is defined as the ratio of nominal GDP divided by real GDP.) The difference between the 1870 to 1940 period and the 1950 to 1995 period is pronounced.

SOURCE: Constructed by the authors from Commerce Department data for 1929–1995. Data for 1869–1928 were kindly provided by Professor Christina Romer.

The stubbornness of the Great Depression shook almost everyone's faith in the ability of the economy to correct itself. In Cambridge, England, this questioning attitude led John Maynard Keynes, one of the world's most reknowned economists, to write *The General Theory of Employment, Interest, and Money* (1936). Probably the most important economics book of the 20th century, it carried a rather revolutionary message. Keynes rejected the idea that the economy always gravitated toward high levels of employment, asserting instead that—if a pessimistic outlook led business firms and consumers to curtail their spending plans—the economy might be condemned to years of stagnation.

While this doleful prognosis sounded all too realistic at the time, Keynes closed his book on a hopeful note. For he showed how government actions might prod the economy out of its depressed state. The lessons he taught the world then are among the lessons we will be learning in Parts VI and VII (along with many qualifications that economists have learned since). These lessons show how governments can manage their economies so that recessions will not turn into depressions and depressions will not last as long as the Great Depression.

While Keynes was working on *The General Theory*, he wrote his friend George Bernard Shaw that, "I believe myself to be writing a book on economic theory which will largely revolutionize . . . the way the world thinks about economic problems." In many ways he was right, though parts of the Keynesian message remain controversial to this day.

LIFE IN "HOOVERVILLE"

[illegible] the worst years of the [illegible] Depression, unemployed [illegible] often congregated in shan[illegible] on the outskirts of many major cities. Conditions in these slums were deplorable. With a heavy dose of irony, these communities were known as "Hoovervilles," in honor of the president of the United States who preached rugged individualism. A contemporary observer described a Hooverville in New York City as follows:

> It was a fairly popular "development" made up of a hundred or so dwellings, each the size of a dog house or chickencoop, often constructed with much ingenuity out of wooden boxes, metal cans, strips of cardboard or old tar paper. Here human beings lived on the margin of civilization by foraging for garbage, junk, and waste lumber. I found some splitting or sawing wood with dull tools to make fires; others were picking through heaps of rubbish they had gathered before their doorways or cooking over open fires or battered [illegible]. Still others spent their days [illegible] their rent-free homes, [illegible] them sometimes fairly solid [illegible]-proof. . . . Most of them, [illegible] to the police, lived by begging or trading in junk; when all else failed they ate at the soup kitchens or public canteens. They were of all sorts, young and old, some of them rough-looking and suspicious of strangers. They lived in fear of being forcibly removed by the authorities, though the neighborhood people in many cases helped them and the police tolerated them for the time being.

SOURCE: Mathew Josephson, *Infidel in the Temple* (New York: Knopf, 1967), pp. 82–83.



FROM WORLD WAR II TO 1973

The Great Depression finally ended when the country mobilized for war in the early 1940s. With government spending at extraordinarily high levels, the economy boomed, and the unemployment rate fell as low as 1.2 percent during the war.

Wartime spending of this magnitude usually leads to inflation, but much of the potential inflation during World War II was contained by price controls. With prices held below the levels at which quantity supplied equaled quantity demanded, many goods had to be rationed, and shortages of consumer goods were common. All of this ended with a burst of inflation when controls were lifted after the war.

The period from the end of the war until the early 1960s was one of strong growth marred by several short recessions. Moderate but persistent inflation also became a fact of life. When the economy emerged from recession in 1961, it

JOHN MAYNARD KEYNES (1883-1946)

John Maynard Keynes was something of a child prodigy. After an outstanding scholastic career at Eton and Cambridge, Keynes took the civil service examination. Ironically, his second-place score was not good enough to land him the position he wanted—in the Treasury. Some years later, reflecting on the fact that his lowest score on the exam was in the economics section, he suggested with characteristic immodesty that, "The examiners presumably knew less than I did." He was probably right.

During World War I, Keynes was called to the Treasury to assist in planning the financial aspects of the war. There his unique combination of daring and intellect quickly established him as a dominant figure. At the war's end, he represented the British Treasury at the peace conference in Versailles. The conference was a turning point in Keynes's life, though it was one of his few failures. He tried unsuccessfully to persuade the Allies to take a less punitive attitude toward the vanquished Germans, and then stormed out of the conference to write his *Economic Consequences of the Peace* (1919), which created a furor. In that work, Keynes argued that the Germans could never meet the harsh economic terms of the treaty, and that its viciousness posed the threat of continued instability and perhaps another war in Europe.

No longer welcome in government, Keynes returned to Cambridge and to his circle of literary and artistic friends in London's Bloomsbury district—a remarkable group that included Virginia Woolf, Lytton Strachey, and E. M. Forster. In 1925, he married the beautiful ballerina Lydia Lopokova, who gave up her stage career for him (though she later acted in a theater that Keynes himself established).

Between the wars, Keynes devoted himself to making money, to economic theory, and to political economy. He managed to make both himself and King's College rich by speculating in international currencies and commodities—allegedly by studying the newspapers while still in bed each morning! In 1936, he published his masterpiece, *The General Theory of Employment, Interest, and Money,* on which much of modern macroeconomics is based.

A heart attack in 1937 reduced Keynes's activities, but he returned to the Treasury during World War II to conduct several delicate financial negotiations with the Americans. Then, as the capstone to a truly remarkable career, he represented Great Britain—and by all accounts dominated the proceedings—at the 1944 conference in Bretton Woods, New Hampshire, that established an international financial system that served the Western world for 27 years. (See Chapter 35.)

He died at home of a heart attack as Lord Keynes, Baron of Tilton, a man who had achieved almost everything that he sought, and who had only one regret: he wished he had drunk more champagne.



entered a period of unprecedented—and noninflationary—growth which was credited widely to the success of what came to be called "The New Economics," a term the media created for the economic policies that Keynes had prescribed in the 1930s. For a while, it looked as if we could avoid both unemployment and inflation. But the optimistic verdicts were premature on both counts.

Inflation came first, beginning about 1966. Its major cause was high levels of wartime spending, as it had been so many times in the past. This time it was the Vietnam war. Unemployment followed when the economy ground to a halt in 1969. Despite a short and mild recession, inflation continued at 5 to 6 percent a year.

Faced with persistent inflation, President Richard Nixon stunned the nation by instituting wage and price controls in 1971, the first time this had ever been done in peacetime. The controls program, which will be discussed in Chapter 33, held inflation in check for a while. But inflation worsened dramatically in 1973, mainly because of an explosion in food prices caused by poor harvests around the world.

THE GREAT STAGFLATION, 1973–1980

Then things began to get much worse, not only for the United States, but for all oil-importing nations. A 1973 war between Israel and the Arab nations led to a quadrupling of the price of oil by the Organization of Petroleum Exporting Countries (OPEC). At the same time, continued poor harvests in many parts of the globe pushed world food prices higher. Prices of other raw materials also skyrocketed. Naturally, higher costs of fuel and other materials soon were reflected in the prices of manufactured goods.

By unhappy coincidence, these events came just as wage and price controls were being lifted. Just as had happened after World War II, the elimination of controls led to a temporary acceleration of inflation as prices that had been held artificially below equilibrium levels were allowed to rise. For all of these reasons, the inflation rate in the United States soared to above 12 percent during 1974.

Meanwhile, the U.S. economy was slipping into what was, up to then, its longest and most severe recession since the 1930s. Real GDP fell between late 1973 and early 1975, and the unemployment rate rose to nearly 9 percent. With both inflation and unemployment unusually virulent in 1974 and 1975, a new term—**stagflation**—was coined to refer to the simultaneous occurrence of economic *stag*nation and rapid in*flation*.

STAGFLATION is inflation that occurs while the economy is growing slowly ("stagnating") or having a recession.

Thanks partly to government actions, but mostly to natural economic forces, a sustained recovery from recession began in 1975. Inflation tumbled rapidly as the adjustment to the end of price controls ended and food and energy prices stopped soaring. The severity of the recession also put a brake on inflation, just as it had in the past. In total, the inflation rate tumbled from over 12 percent back down to the 6 percent range.

But the price of oil soared again in 1979 following a revolution in Iran, bringing stagflation back. This time, inflation hit the astonishing rate of 16 percent during the first half of 1980 and the government clamped on credit controls. The controls broke the back of the economy and output fell at an extraordinarily rapid pace. But credit controls remained in effect for only a few months, and by late 1980 recovery was under way.

REAGANOMICS AND ITS AFTERMATH

When President Ronald Reagan assumed office in January 1981, the U.S. economy was showing signs of reviving, but the inflation rate seemed stuck near 10 percent. The new president promised to change things with a package of policies called "supply-side economics."[7]

At first, things did change dramatically—but not in the way President Reagan wanted. While inflation fell remarkably to only about 4 percent in 1982, the lowest rate in a decade, the economy slumped into its worst recession since the Great Depression. When the 1981–1982 recession hit bottom, the unemployment rate was approaching 11 percent, the financial markets were in disarray, and the word *depression* had reentered the American vocabulary. The U.S. government also had a huge budget deficit, bigger than anyone had dreamed possible only a few years before. This problem, as we shall see in subsequent chapters, is still with us.

[7]Supply-side economics is discussed further in Chapter 28.

However, the recovery that began in the winter of 1982–1983 proved to be one of the most vigorous and long-lasting in our history. Unemployment fell more or less steadily for about 6 years, eventually dropping below 5-1/2 percent. Meanwhile, inflation remained tame. All of this provided an ideal economic platform on which George Bush ran to succeed Reagan—and to continue Reagan's policies.

Unfortunately for President Bush, the good times did not keep rolling. Shortly after he took office, inflation began to accelerate a bit, economic growth began to sputter, and Congress enacted a deficit-reduction package (including a tax increase) not entirely to the president's liking. Then, in 1990–1991, the U.S. economy slumped into another recession—precipitated, according to some observers, by yet another spike in oil prices before the Persian Gulf war.

While the 1990–1991 recession was relatively mild, the economy had not recovered by the time of the 1992 election. In fact, the growth rate during George Bush's presidency was the weakest of any 4-year period since World War II. This fact was not lost on candidate Bill Clinton, who hammered away at the lackluster economic performance of the period from 1989 to 1992. Most observers believe that the weak economy was the main factor behind George Bush's electoral defeat.

CLINTONOMICS: DEFICIT REDUCTION AND THE RESUMPTION OF GROWTH[8]

President Clinton ran on a detailed economic platform that concentrated on two objectives: spurring economic growth and increasing public investment. But, even before his inauguration in January 1993, the yawning budget deficit forced the new president to deemphasize new spending and concentrate on deficit reduction. A politically contentious package of substantial tax increases and major spending cuts passed the Congress in August 1993 without a single Republican vote. Since then, deficit reduction has been the major economic issue on the American political landscape.

Whether by cause or coincidence, the national economy improved dramatically in the first 2 years of the Clinton presidency. Business perked up, unemployment fell rapidly, and inflation remained remarkably well contained—in the 3 percent range. By the spring of 1996, macroeconomic conditions were the best in a generation: The economic expansion was more than 5 years old, unemployment was about 5-1/2 percent, and inflation was still below 3 percent.

THE PROBLEM OF MACROECONOMIC STABILIZATION: A SNEAK PREVIEW

STABILIZATION POLICY is the name given to government programs designed to prevent or shorten recessions and to counteract inflation (that is, to *stabilize* prices).

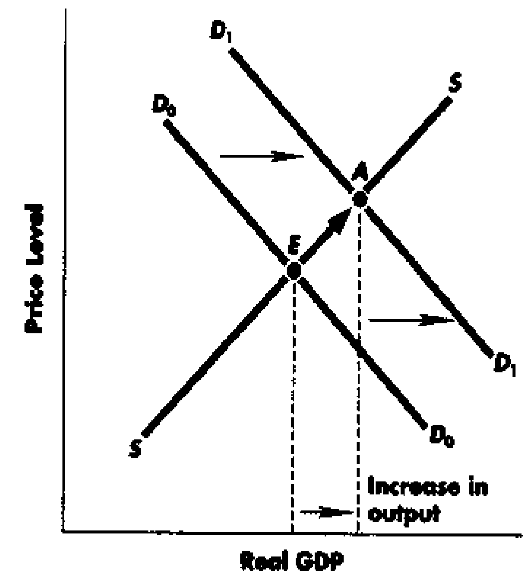
This brief look at the historical record shows that our economy has not generally produced steady growth without inflation. Rather, it has been buffeted by periodic bouts of unemployment or inflation, and sometimes has been plagued by both. There was also a hint that government policies may have had something to do with this performance. Let us now expand briefly upon this hint—a subject that will be developed further in subsequent chapters.

We can provide a preliminary analysis of **stabilization policy**, the name given to government programs designed to prevent or shorten recessions and to coun-

[8]One of the authors of this book was a member of President Clinton's original Council of Economic Advisers.

FIGURE 22-5 STABILIZATION POLICY TO FIGHT UNEMPLOYMENT

This diagram duplicates Figure 22-1(b), but here we assume that point E—the intersection of the demand curve D_0D_0 and the supply curve SS—corresponds to high unemployment. With the kind of policy tools that we will study in later chapters, the government could shift the aggregate demand curve outward to D_1D_1. This would raise output and lower unemployment.

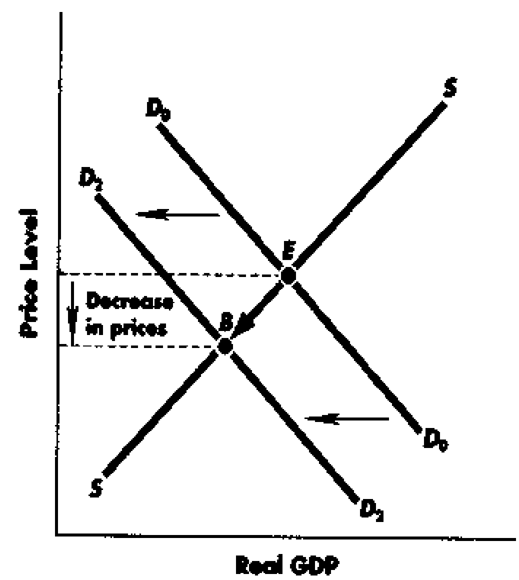


teract inflation, by using the basic tools of aggregate supply and aggregate demand analysis. To facilitate this, we have reproduced as Figures 22-5 and 22-6 two diagrams found earlier in this chapter, but we now give them slightly different interpretations.

Figure 22-5 offers a simplified view of government policy to fight unemployment. Suppose that, in the absence of government intervention, the economy would reach an equilibrium at point E, where the demand curve D_0D_0 crosses the supply curve SS. Now if the output corresponding to point E is so low that many workers are unemployed, *the government can reduce unemployment by increasing aggregate demand.* Chapter 28 will consider in detail how the

FIGURE 22-6 STABILIZATION POLICY TO FIGHT INFLATION

This diagram duplicates Figure 22-2, but here we assume that point E—the equilibrium that the economy would attain without government intervention—represents high inflation (that is, the price level corresponding to point E is far above last year's price level). By using its policy instruments to shift the aggregate demand curve inward to D_2D_2, the government can keep this year's price level lower than it would otherwise have been; in other words, the government can reduce inflation.



government might do this. In the diagram, such an action would shift the demand curve to D_1D_1, causing equilibrium to move to point A. In general:

Recessions and unemployment are often caused by insufficient aggregate demand. When this occurs, government policies that successfully augment demand—such as increases in government spending—can be an effective way to increase output and reduce unemployment.

The opposite type of demand management is called for when inflation is the main macroeconomic problem. Figure 22-6 illustrates this case. Here again, point E, the intersection of the demand curve D_0D_0 and the supply curve SS, is the equilibrium that would be reached in the absence of government policy. But now we suppose that the price level corresponding to point E is considered "too high," meaning that the *change* in the price level from the previous period to this one would be too rapid if the economy moved to point E. A government program that reduces demand from D_0D_0 to D_2D_2 (for example, a reduction in government spending) can keep prices down and thereby reduce inflation. Thus:

Inflation is frequently caused by aggregate demand racing ahead too fast. When this is the case, government policies that reduce aggregate demand can be effective anti-inflationary devices.

This, in brief, summarizes the intent of stabilization policy. When aggregate demand fluctuations are the source of economic instability, the government can limit both recessions and inflations by pushing aggregate demand ahead when it would otherwise lag, and restraining it when it would otherwise grow too quickly.

Does it work? Can the government actually stabilize the economy? That is a matter of some debate, a debate we will take up in Part VII. But a look back at Figures 22-3 and 22-4 may be enlightening right now. First, cover the portions of Figures 22-3 and 22-4 that deal with the period after 1940, the portions from the shaded area rightward in each figure. The picture that emerges for the 1870 to 1940 period is of an economy with frequent and sometimes quite pronounced fluctuations.

Now do the reverse. Cover the data before 1950 and look only at the postwar period. There is, indeed, a difference. Instances of negative real GDP growth are less common and business fluctuations look less severe. While perfection has not been achieved, things do look much better. When we turn to inflation, however, things look rather worse. Gone are the periods of deflation and price stability that occurred before World War II. Prices now seem only to rise.

This quick tour through the data suggests that something has changed. The U.S. economy behaved differently from 1950 to 1995 than it did from 1870 to 1940. Although there is controversy on this point, many economists attribute this shift in the economy's behavior to lessons the government has learned about managing the economy—lessons we will be learning in Part VII.

When you look at the prewar data, you are looking at an unmanaged economy that went through booms and recessions for "natural" economic reasons. The government did little about either. When you examine the postwar data, on the other hand, you are looking at an economy that has been increasingly managed by government policy—sometimes successfully and sometimes unsuccessfully. While the recessions are less severe, this improvement has come at a cost: The economy appears to be more inflation-prone than it was in the more distant

past. These two changes in our economy may be connected. But, to understand why, we will have to provide some relevant economic theory. That is the task of the rest of Part VI.

SUMMARY

1. *Microeconomics* studies the decisions of individuals and firms, how these decisions interact, and how they influence the allocation of society's resources and the distribution of income. *Macroeconomics* looks at the behavior of entire economies and studies the pressing social problems of inflation and unemployment.
2. While they focus on different subjects, microeconomics and macroeconomics rely on virtually identical tools. Both use the supply and demand analysis introduced in Chapter 4.
3. Macroeconomic models use abstract concepts like "the price level" and "domestic product" that are derived by amalgamating many different markets into one. This process is known as **aggregation**; it should not be taken literally but rather viewed as a useful approximation.
4. The best specific measure of the nation's economic output is **gross domestic product (GDP)**, which is obtained by adding up the money values of all **final goods and services** produced in a given year. These outputs can be evaluated year by year at current market prices (to get **nominal GDP**) or at some common set of prices (to get **real GDP**). Neither **intermediate goods** nor transactions that take place outside organized markets are included in GDP.
5. The GDP is meant to be a measure of the *production* of the economy, not of the increase in its *well-being*. For example, the GDP places no value on housework and other do-it-yourself activities or on leisure time. On the other hand, even commodities that might be considered as "bads" rather than "goods" are counted in the GDP (for example, activities that harm the environment).
6. America's economic history is one of growth punctuated by periodic **recessions**, that is, periods in which real GDP declined. While the distant past included some periods of falling prices (**deflation**), more recent history shows only rising prices (**inflation**).
7. The Great Depression of the 1930s was the worst in our country's history. It had profound effects both on our nation and on countries throughout the world, and it led also to a revolution in economic thinking, thanks to the work of John Maynard Keynes.
8. From World War II to the early 1970s, the American economy exhibited much steadier growth than it had shown in the past. Many observers attributed this to the implementation of the **stabilization policies** that Keynes suggested. At the same time, however, the price level seems only to rise, never to fall, in the modern economy. The economy seems to have become more "inflation-prone."
9. Since 1973, the U.S. economy has suffered through several serious recessions. Between 1973 and about 1981, inflation was also unusually virulent. This unhappy combination of economic stagnation with rapid inflation was nicknamed "**stagflation.**" Since 1982, however, inflation has been low and fairly steady.
10. One major cause of inflation is that **aggregate demand** may grow more quickly than **aggregate supply**. In such a case, a government policy that reduces aggregate demand may be able to stem the inflation.
11. Similarly, recessions often occur because aggregate demand grows too slowly. In this case, a government policy that stimulates demand may be an effective way to fight the recession.

KEY TERMS

Microeconomics
Macroeconomics
Aggregation
Aggregate demand curve
Aggregate supply curve
Inflation
Recession
Gross domestic product (GDP)
Nominal GDP
Real GDP
Final goods and services
Intermediate good
Deflation
Stagflation
Stabilization policy

QUESTIONS FOR REVIEW

1. Which of the following problems are likely to be studied by a microeconomist and which by a macroeconomist?
   a. The growth of Microsoft Corporation
   b. Why unemployment in the United States fell in 1993 and 1994
   c. Why Japan's economy grew so much faster than the United States' economy for several decades, but has failed to do so thus far in the 1990s
   d. Why university budgets have come under stress in recent years
2. You probably use "aggregates" frequently in everyday discussions. Try to think of some examples. (Here is one: Have you ever said, "The students at this college generally think . . ."? What, precisely, did you mean?)
3. Use an aggregate supply and demand diagram to study what would happen to an economy in which the aggregate supply curve never moved while the aggregate demand curve shifted outward year after year.
4. Try asking a friend who has not studied economics in which year he or she thinks prices were higher: 1870 or 1900? 1920 or 1940? (In both cases, prices were higher in the earlier year.) Most people your age think that prices have always risen. Why do you think they have this opinion?
5. Which of the following transactions are included in gross domestic product, and by how much does each raise GDP?
   a. Smith pays a carpenter $15,000 to build a garage.
   b. Smith purchases $3,000 worth of materials and builds himself a garage, which is worth $15,000.
   c. Smith goes to the woods, cuts down a tree, and uses the wood to build himself a garage that is worth $15,000.
   d. The Jones family sells its old house to the Reynolds family for $100,000. The Joneses then buy a newly constructed house from a builder for $150,000.
   e. You purchase a used computer from a friend for $200.
   f. Your university purchases a new mainframe computer from IBM, paying $200,000.
   g. You win $1,000 in an Atlantic City casino.
   h. You make $1,000 in the stock market.
   i. You sell a used economics textbook to your college bookstore for $25.
   j. You buy a new economics textbook from your college bookstore for $50.
6. Give some reasons why gross domestic product is not a suitable measure of the well-being of the nation. (Have you noticed newspaper accounts in which journalists seem to use GDP for this purpose?)



CHAPTER 23

Unemployment and Inflation: The Twin Evils of Macroeconomics

When men are employed, they are best contented.
Benjamin Franklin

Inflation is repudiation.
Calvin Coolidge

Among the many trials faced by Odysseus, the hero of Homer's *Odyssey*, one of the most difficult was to steer his fragile boat through a narrow strait. On one side lay the rock of the monster Scylla, which threatened to break his craft into pieces; on the other was the menacing whirlpool of Charybdis. The makers of national economic policy face a similarly difficult task in trying to chart a middle course between the Scylla of unemployment and the Charybdis of inflation. If they steer the economy far from the rocks of unemployment, they run the risk of being swept up in the swift currents of inflation. But if they maintain a safe distance from inflation, they may smash against the rocks of unemployment.

In Parts VI and VII, we will explain how economic planners attempt to strike a balance between high employment and low inflation, why these goals cannot be attained with machine-like precision, and why improvement on one front generally spells deterioration on the other. We will pay a great deal of attention to both the *causes* of and *cures* for inflation and unemployment.

But before getting involved in such weighty issues of theory and policy, we pause in this chapter to take a close look at the twin evils themselves. Why does a rise in unemployment cause such social distress? Why is inflation so loudly deplored? Can we measure the costs of unemployment and inflation? The answers to some of these questions may seem obvious at first. But we will see that there is more to them than meets the eye.

The chapter is divided into two parts. The first deals with unemployment. After discussing the human and economic costs of high unemployment, we explain how government statisticians measure unemployment. Then we consider how the elusive concept of "full employment" can be defined. We conclude by investigating our country's system of unemployment insurance.

The second part of the chapter is devoted to inflation. We begin by dispelling some persistent myths about inflation. But the costs of inflation are not all mythical. For example, inflation capriciously redistributes income and wealth from one group of people to another, and certain laws make inflation impose economic damage that could be avoided if the laws were written differently. This last cost stems from failure to understand one of the **Ideas for Beyond the Final Exam:** the effect of inflation on interest rates. Finally, we contrast the costs of low versus high inflation and dispel yet another myth: that low inflation inevitably gives way to high inflation.

An appendix at the end of this chapter explains how inflation is measured.

THE COSTS OF UNEMPLOYMENT

The human costs of unemployment are probably sufficiently obvious. Years ago, loss of a job meant not only enforced idleness and a catastrophic drop in income, it often led to hunger, cold, ill health—even death. This is the way one unemployed worker during the Great Depression described his family's plight in a mournful letter to the governor of Pennsylvania:

> I have been out of work for over a year and a half. Am back almost thirteen months and the landlord says if I don't pay up before the 1 of 1932 out I must go, and where am I to go in the cold winter with my children? If you can help me please for God's sake and the children's sakes and like please do what you can and send me some help, will you, I cannot find any work. . . . Thanksgiving dinner was black coffee and bread and was very glad to get it. My wife is in the hospital now. We have no shoes to were [sic]; no clothes hardly. Oh what will I do I sure will thank you.[1]

Nowadays, unemployment does not hold quite such terrors for most families, although its consequences remain dire enough. Our system of unemployment insurance (discussed later in the chapter) has taken out part of the sting of unemployment, and there are other social welfare programs to support the incomes of the poor. Yet most families still suffer painful losses of income and, often, severe noneconomic consequences when a breadwinner becomes unemployed.

Even families that are well-protected by unemployment compensation suffer when joblessness strikes. Ours is a work-oriented society. A man's place has always been in the office or shop, and lately this has become true for many women, as well. A worker forced into idleness by a recession endures a psychological cost that is no less real for our inability to measure it. Martin Luther King put it graphically: "In our society, it is murder, psychologically, to deprive a man of a job. . . . You are in substance saying to that man that he has no right to exist."[2] High unemployment has been linked to psychological and even physical disorders, divorces, suicides, and crime.

Furthermore, accumulated work experience is a valuable asset. When forced into idleness, workers not only cease accumulating experience, but lengthy peri-

---

[1]From *Brother, Can You Spare a Dime? The Great Depression 1929–1933*, by Milton Meltzer, p. 103. Copyright 1969 by Milton Meltzer. Reprinted by permission of Alfred A. Knopf, Inc.

[2]Quoted in Coretta Scott King (ed.), *The Words of Martin Luther King* (New York: Newmarket Press, 1983), p. 45.

ods of unemployment may make them "rusty," and thus less productive when they are reemployed. Short periods of unemployment exact different kinds of costs. A record of steady employment is important in applying for a new job, and a worker who has frequently been laid off will lack this record.

The **UNEMPLOYMENT RATE** is the number of unemployed people, expressed as a percentage of the labor force. The labor force is the number of people holding or seeking jobs.

It is important to realize that these costs, whether large or small in total, are distributed most unevenly across the population. In 1995, for example, the **unemployment rate** among all workers averaged 5.6 percent. But, as Figure 23-1 shows, 10.4 percent of black workers were unemployed, as were 8 percent of women who maintained families. For teenagers, the situation was worse still, with unemployment at 17.3 percent, and that of black teenagers about 36 percent. Married men had the lowest rate—about 3.3 percent. Overall unemployment varies from year to year, but these relationships are typical:

In good times and bad, married men suffer the least unemployment and teenagers suffer the most; nonwhites are unemployed much more often than whites; blue-collar workers have above-average rates of unemployment; well-educated people have below-average unemployment rates.

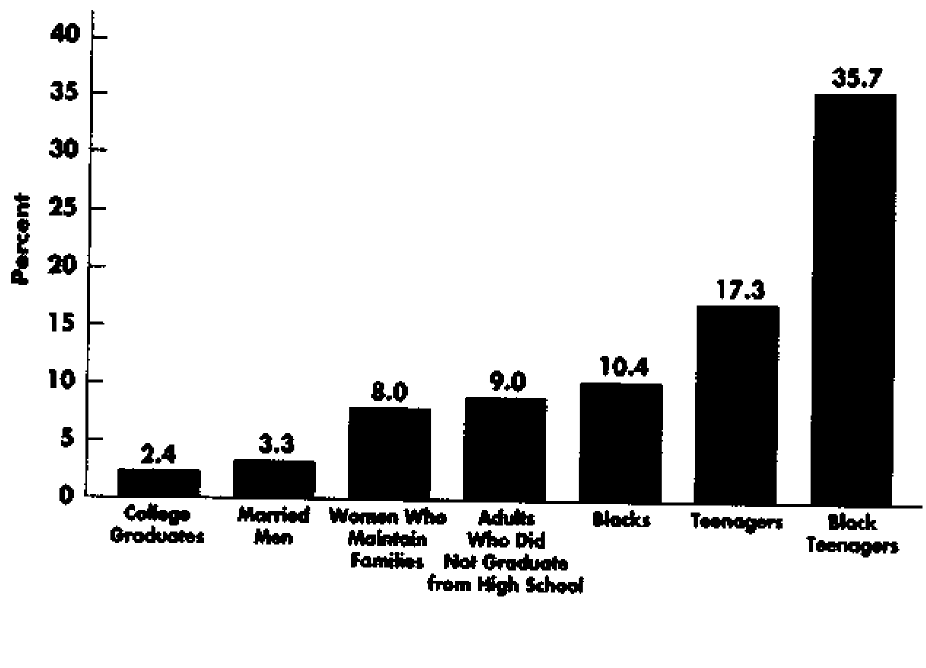
THE ECONOMIC COSTS OF HIGH UNEMPLOYMENT

Some of the human costs of high unemployment are, as we just noted, intangible. But others can be translated directly into dollars and cents because:

When the economy does not generate enough jobs to employ all those who are willing to work, a valuable resource is lost. Potential goods and services that might have been enjoyed by consumers are lost forever. This is the real economic cost of high unemployment.

And these costs are hardly negligible. Table 23-1 summarizes the idleness of workers and machines, and the resulting loss of national output, for some of the

FIGURE 23-1 UNEMPLOYMENT RATES FOR SELECTED GROUPS, 1995



This figure shows that, in 1995, unemployment rates for specific demographic groups varied widely. Similar patterns hold in most years.

SOURCE: Bureau of Labor Statistics.

TABLE 23-1 THE ECONOMIC COSTS OF HIGH UNEMPLOYMENT

Year	Civilian Unemployment Rate (percent)	Capacity Utilization Rate (percent)	Percentage of Real GDP Lost Due to Idle Resources
1958	6.8	75.0	4.4
1961	6.7	77.3	3.3
1975	8.5	72.3	4.6
1982	9.7	70.3	7.4
1992	7.5	79.5	1.6
1995	5.6	82.9	−0.1

SOURCES: Bureau of Labor Statistics; Federal Reserve System; and the authors' calculations.

years of lowest economic activity in recent decades. The second column lists the civilian unemployment rate, and thus measures unused labor resources. The third lists the percentage of industrial capacity that U.S. manufacturers were actually using, and thus indicates the extent of unused plant and equipment. The fourth column is an estimate of how much more output (real GDP) could have been produced if these labor and capital resources had been fully employed. For comparison, the bottom line shows the situation in 1995, a year of approximately full utilization of resources. We see that unemployment has cost the people of the United States as much as a 7-1/2 percent reduction in their real incomes.

While Table 23-1 shows extreme examples, inability to utilize all of the nation's available resources has been a recurrent problem for our economy since 1973. The orange line in Figure 23-2 shows actual real GDP in the United States from 1954 to 1995, while the black line shows the real GDP we *could have* produced if "full employment" had been maintained. This last statement defines a concept called **potential GDP**. The graph shows that actual GDP has at times fallen short of potential GDP of the economy.

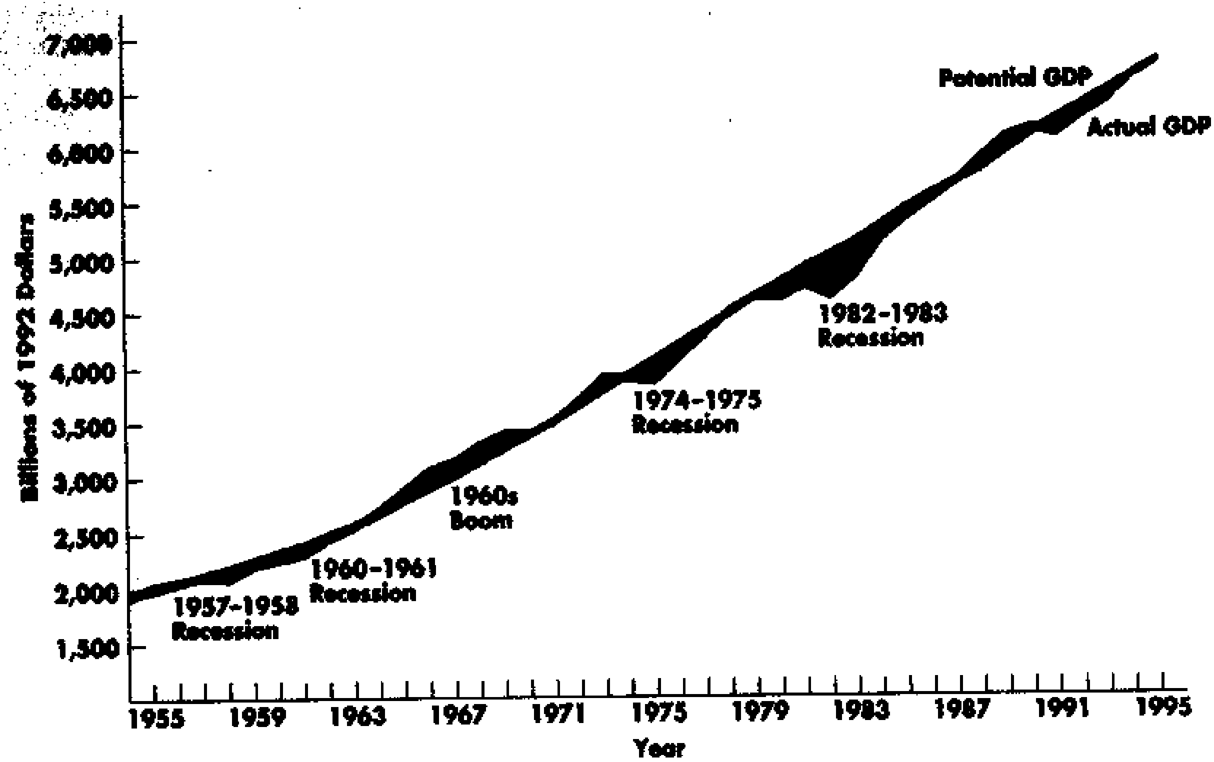
POTENTIAL (GDP) is the real GDP that the economy would produce if its labor and other resources were fully employed.

It *is* possible to push employment beyond its normal full-employment level. This occurs whenever the unemployment rate dips below the "full-employment unemployment rate." Consequently, it *is* possible for actual GDP to exceed potential GDP. Figure 23-2 shows several instances where this happened, shaded in blue. But it also shows that actual GDP has fallen short of potential GDP often since 1973—sometimes by huge amounts. In fact:

A conservative estimate of the cumulative gap between actual and potential GDP over the years 1974 to 1995 (all evaluated in 1992 prices) is approximately $1,630 billion. At 1995 levels, this loss in output as a result of unemployment would be about 3 months' worth of production. And there is no way to redeem these losses. The labor wasted in 1991 cannot be utilized in 1996.

Those who argue that unemployment is nothing to worry about today because of unemployment insurance, or because unemployment is concentrated among certain kinds of workers (such as teenagers), or because many unemployed workers become reemployed within a few weeks, should ponder Figure 23-2. Is the loss of this much output really no cause for worry? Would these optimists

FIGURE 23-2 ACTUAL AND POTENTIAL GDP IN THE UNITED STATES, 1954–1995



This chart compares the growth of actual GDP (orange line) with that of potential GDP (black line). There have been two lengthy periods during which real GDP remained below its potential (1957 to 1963 and 1974 to 1987), but only one lengthy period during which GDP remained above its potential (1965 to 1970). The large shortfalls of GDP from its potential in the early 1980s stand out.

SOURCE: U.S. Department of Commerce and the authors' calculations.

react the same way if the government collected a fraction of the output of every factory in America and dumped it into the sea? Waste is waste no matter who ultimately pays the cost.

COUNTING THE UNEMPLOYED: THE OFFICIAL STATISTICS

We have been using figures on unemployment without considering where they come from or how accurate they are. The basic data come from a monthly survey of over 50,000 households conducted for the Bureau of Labor Statistics (BLS). The census-taker asks several questions about the employment status of each member of the household and, on the basis of the answers, classifies each person as *employed, unemployed,* or *not in the labor force.*

The first category is the simplest to define. It includes everybody currently working at a job, including part-time workers. Although some part-timers work less than a full week by choice, others do so only because they cannot find suitable full-time jobs. Nevertheless, these workers are counted as employed, even though many would consider them "underemployed."

The second category is a bit trickier. For those not currently working, the BLS first determines whether they are temporarily laid off from a job to which they

expect to return. If so, they are counted as unemployed. The remaining workers are asked whether they actively sought work during the previous 4 weeks. If they did, they are also counted as unemployed. But if they did not, they are classified as *out of the labor force:* that is, since they failed to look for a job, they are not considered unemployed.

This seems a reasonable way to draw the distinction—after all, we do not want to count college students who work during the summer months as unemployed between September and May. Yet, there is a problem: Research has shown that many unemployed workers give up looking for jobs after a time. These so-called **discouraged workers** are victims of poor job prospects, just like the officially unemployed. But, ironically, when they give up hope, measured unemployment declines! Some critics have therefore argued that an estimate should be made of the number of discouraged workers and that these people should be added to the rolls of the unemployed. The BLS estimated that about 400,000 people fell into this category in 1995.

A **DISCOURAGED WORKER** is an unemployed person who gives up looking for work and is therefore no longer counted as part of the labor force.

Involuntary part-time work, loss of overtime or shortened work hours, and discouraged workers are all examples of "hidden" or "disguised" unemployment. Those who are concerned about these phenomena argue that we should include them in the official unemployment rate because, if we do not, the magnitude of the problem will be underestimated. Others, however, argue that measured unemployment overestimates the problem because, to count as unemployed, a person need only *claim* to be looking for a job, even if he or she is not really interested in finding one.

TYPES OF UNEMPLOYMENT

Providing jobs for those willing to work is one principal goal of macroeconomic policy. How are we to define this goal? One clearly *incorrect* answer would be "a zero measured unemployment rate." Ours is a dynamic, highly mobile economy. Households move from one state to another. Individuals quit jobs to seek better positions or retool for more attractive occupations. These and other phenomena produce some minimal amount of unemployment—people who are literally *between* jobs. Economists call this the level of **frictional unemployment.**

FRICTIONAL UNEMPLOYMENT is unemployment that is due to normal turnover in the labor market. It includes people who are temporarily bewteen jobs because they are moving or changing occupations, or for similar reasons.

The critical distinguishing feature of frictional unemployment is that it is short-lived. A frictionally unemployed person has every reason to expect to find a new job soon. People tend to think of frictional unemployment as irreducible, but that is not true. During World War II, for example, unemployment in this country fell below 2 percent—substantially below the frictional level.

Frictional unemployment is irreducible only in the sense that—under normal circumstances—it is socially undesirable to do so. Geographic and occupational mobility play important roles in our market economy, enabling people to search for better jobs. Similarly, waste is avoided by allowing inefficient firms, or firms producing items no longer in demand, to be replaced by new firms. Inhibiting either of these phenomena would hamper the workings of the market economy. But, if these adjustment mechanisms are allowed to operate, there will always be some temporarily unemployed workers looking for jobs, and there will always be some firms with unfilled positions looking for workers. This is the genesis of frictional unemployment.

STRUCTURAL UNEMPLOYMENT refers to workers who have lost their jobs because they have been displaced by automation, because their skills are no longer in demand, or for similar reasons.

A second type of unemployment is often difficult to distinguish from frictional unemployment, but it has very different implications. **Structural unemployment** arises when jobs are eliminated by changes in the structure of the economy, such

as automation or permanent changes in demand. The crucial difference between frictional and structural unemployment is that, unlike frictionally unemployed workers, structurally unemployed workers cannot realistically be considered "between jobs." Instead, they may find their skills and experience unwanted in the changing economy in which they live. They are thus faced with either prolonged periods of unemployment or the necessity of making major changes in their occupations. For older workers, learning a new occupation may be nearly impossible.

CYCLICAL UNEMPLOYMENT is the portion of unemployment that is attributable to a decline in the economy's total production. Cyclical unemployment rises during recessions and falls as prosperity is restored.

The remaining type of unemployment, **cyclical unemployment**, will occupy most of our attention. Cyclical unemployment rises when the level of economic activity declines, as it does in a recession. Thus, when economists speak of maintaining "full employment," they do not mean achieving zero measured unemployment, but rather limiting unemployment to its frictional and structural components. A key question, therefore, is: How much measured unemployment is that?

HOW MUCH EMPLOYMENT IS "FULL EMPLOYMENT"?

President John F. Kennedy first committed the federal government to a specific numerical goal for unemployment. Looking at experience in the prosperous early 1950s, he picked a 4 percent target. But the 4 percent goal was rejected as outmoded during the 1970s—for three major reasons.

First, some economists argued that the 4 percent target had to be raised because the composition of the labor force had changed. Specifically, there were many more young workers in the 1970s than in the 1950s, and teenagers always have higher rates of unemployment than adults.

Second, they suggested, the increased generosity of unemployment compensation (which is discussed just below) had reduced the incentive to get off the unemployment rolls. The logic here is simple. Since unemployment insurance reduces the income gap between those who work and those who collect unemployment benefits, it must therefore dull the incentive to look for and then accept a job. And the more generous the benefits, the weaker the incentive.

A third reason stems from the short-run *trade-off between inflation and unemployment,* a notion which was mentioned in Chapter 1 and will be studied in depth in Chapter 33. Economic research conducted in the 1970s suggested that the original 4 percent unemployment target was too low—even for the 1960s—in a very specific sense: that inflation will rise if unemployment remains this low.

While the 4 percent unemployment target was abandoned, no new number was put in its place. Instead, a debate began over exactly how much measured unemployment corresponds to *full employment*—a debate that continues to this day. In the 1980s, some economists argued that full employment came at a measured unemployment rate above 6 percent. Others pointed out that the main factors that had raised the full-employment unemployment rate in the 1970s were reversed in the 1980s. For example, the teenage labor force dwindled, and unemployment benefits went to a smaller percentage of the unemployed.

As is so often the case, actual events helped settle the issue—though not definitively. Measured unemployment fell below 6 percent late in the 1980s and remained there until the end of 1990. And, just as this was happening, inflation started to accelerate. This conjunction of events persuaded many economists that full employment comes at an unemployment rate near 6 percent. Then, in 1994 to 1996, inflation remained stable while unemployment hovered just above 5-1/2

POLICY DEBATE DOES THE MINIMUM WAGE CAUSE UNEMPLOYMENT?

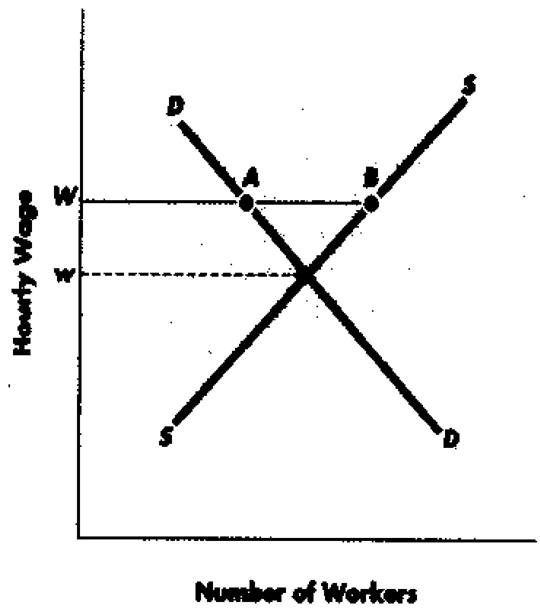
Elementary economic reasoning—summarized in the simple supply-demand diagram below—suggests that setting a minimum wage (*W* in the graph) above the free-market wage (*w* in the graph) must cause unemployment. In the graph, unemployment is the horizontal gap between the quantity of labor supplied (point *B*) and quantity demanded (point *A*) at the minimum wage. Indeed, the conclusion seems so elementary that generations of economists took it for granted. Earlier editions of this book, for example, confidently told students that a higher minimum wage must lead to higher unemployment.

But some surprising economic research published in the 1990s casts serious doubt on this conventional wisdom.* For example, Princeton economists David Card and Alan Krueger compared employment changes at fast-food restaurants in New Jersey and nearby Pennsylvania after New Jersey, but not Pennsylvania, raised its minimum wage in 1992. To their surprise, the New Jersey stores did more net hiring than their Pennsylvania counterparts. Similar results were found for fast-food stores in Texas after the federal minimum wage was raised in 1991, and in California after the statewide minimum wage was increased in 1988.

In none of these cases did a higher minimum wage seem to reduce employment—in contrast to the implications of simple economic theory. Thus, a policy question previously deemed closed now seems to be open: Does the minimum wage cause unemployment?

Resolution of this debate is of more than academic interest. In 1995, the Clinton administration recommended an increase in the federal minimum wage—justifying its request, in part, by the new research suggesting that unemployment would not rise as a result. But the Republican Congress, adhering to the conventional view that a higher minimum wage would raise unemployment, refused to go along.

*See David Card and Alan Krueger, *Myth and Measurement: The New Economics of the Minimum Wage* (Princeton, N.J.: Princeton University Press, 1995).



percent, leading many economists to adopt 5-1/2 percent as a working definition of full employment.

UNEMPLOYMENT INSURANCE: THE INVALUABLE CUSHION

One main reason why America's unemployed workers no longer experience the complete loss of income that devastated so many during the 1930s is our system of *unemployment insurance*—one of the most valuable institutional innovations to emerge from the trauma of the Great Depression.

Each of the 50 states now administers an unemployment insurance program under federal guidelines. While the precise amounts vary, the average weekly benefit check in 1995 was about $187, which amounted to about 47 percent of average earnings. Though a 53 percent drop in earnings still poses serious problems, the importance of this 47 percent income cushion can scarcely be exagger-

ated, especially since it may be supplemented by funds from other welfare programs. Families that are covered by unemployment insurance rarely go hungry or are dispossessed from their homes when they lose their jobs.

Who is eligible to receive these benefits? Precise criteria vary by state, but some stipulations apply quite generally. Only experienced workers qualify; so persons just joining the labor force (such as recent graduates of high schools and colleges) or reentering after prolonged absences (such as women resuming work after years of child rearing) cannot collect benefits. Neither can those who quit their jobs, except under unusual circumstances. Also, benefits end after a stipulated period of time, normally 6 months. For all of these reasons, only about one-third of the 8 million people who were unemployed in an average week in 1994 actually received benefits. Many of the rest received no government support.

The importance of unemployment insurance to the unemployed is obvious. But there are also significant benefits to citizens who never become unemployed. During recessions, many billions of dollars are paid out in unemployment benefits, and since recipients probably spend most of their benefits, unemployment insurance limits the severity of recessions by providing additional purchasing power when and where it is most needed.

The unemployment insurance system is one of several "cushions" that have been built into our economy since 1933 to prevent another Great Depression. By giving money to those who become unemployed, the system helps prop up aggregate demand during recessions.

While the U.S. economy is now probably "depression proof," this should not be a cause for too much rejoicing, for the recession of 1990–1991 amply demonstrated that we are far from "recession proof."

UNEMPLOYMENT INSURANCE AND THE COSTS OF UNEMPLOYMENT

The fact that unemployment insurance and other social welfare programs replace a significant fraction of lost income has led some skeptics to claim that unemployment is no longer a serious problem. But the fact is that:

Unemployment insurance is just what the name says—an *insurance* program. And insurance can never prevent a catastrophe from occurring; it can only *spread the costs* of a catastrophe among many people instead of letting all the costs fall on the shoulders of those few unfortunate souls who it affects directly.

Fire insurance is an example. If your family is covered by fire insurance and your house burns down, you will probably suffer only a small financial loss because the insurance company will pay most of the expenses. But it does not get the money out of thin air. Rather, it must have collected the funds from the many other families who purchased insurance but did not suffer any fire damages. Thus, one family's loss of perhaps $100,000 is covered by the insurance payments of 500 families each paying $200 a year. In this way, the costs of the catastrophe are spread among hundreds of families, and in the process, made much more bearable.

But despite the insurance, the family whose house is destroyed by fire suffers anguish and inconvenience. No insurance policy can eliminate this. Furthermore, society loses a valuable resource—a house. It will take much wood, cement, nails, paint, and labor to replace the burnt-out home. *An insurance policy cannot insure society against losses of real resources.*

The case is precisely the same with insurance against unemployment. All workers and employers pay for the insurance policy by a tax that the government levies on wages and salaries. With the funds so collected, the government compensates the victims of unemployment. Thus, instead of letting the costs of unemployment fall entirely on the minority of workers who lose their jobs:

Our system of payroll taxes and unemployment benefits *spreads* the costs of unemployment over the entire population. But it does not eliminate the basic economic cost.

THE COSTS OF INFLATION

Both the human and economic costs of inflation are less obvious than the costs of unemployment. But this does not necessarily make them any less real, for if one thing is crystal clear about inflation, it is that people do not like it.

Public opinion polls show that, except when inflation is extremely low, it ranks high on people's list of major national problems—generally even ahead of unemployment. Surveys also find that inflation, like unemployment, causes a deterioration in consumers' sense of well-being—it makes people unhappy. Finally, studies of elections suggest that voters penalize the party that occupies the White House when inflation is high.

The fact is beyond dispute: People consider inflation to be something bad. The question is: Why?

INFLATION: THE MYTH AND THE REALITY

At first, the question may seem ridiculous. During times of inflation people keep paying higher prices for the same quantities of goods and services they had before. So more and more income is needed just to maintain the same standard of living. Is it not obvious that this erosion of **purchasing power**—that is, the decline in what money will buy—makes everyone worse off?

The **PURCHASING POWER** of a given sum of money is the volume of goods and services that it will buy.

This would indeed be the case were it not for one very significant fact. The wages that people earn are also prices—prices for labor services. During a period of inflation, wages also rise, and, in fact, the average wage typically rises more or less in step with prices. Thus, contrary to popular myth, workers as a group are not usually victimized by inflation.

The **REAL WAGE RATE** is the wage rate adjusted for inflation. It indicates the volume of goods and services that money wages will buy.

The purchasing power of wages—what is called the **real wage rate**—is not systematically eroded by inflation. Sometimes wages rise faster than prices, and sometimes prices rise faster than wages. The fact is that, in the long run, wages tend to outstrip prices as new capital equipment and innovation increase output per worker.

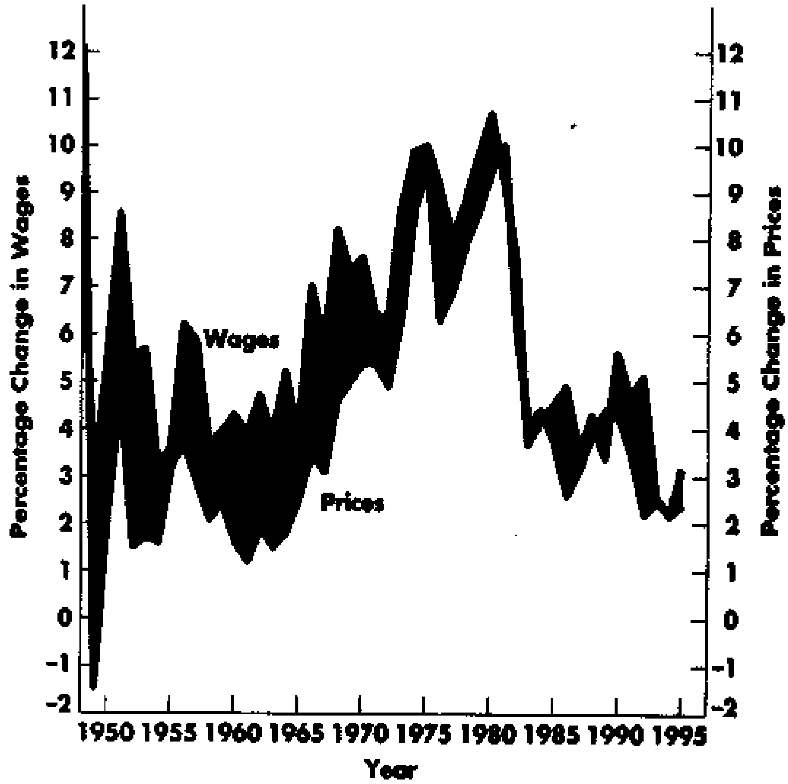
Figure 23-3 illustrates this simple fact. The orange line shows the annual rate of increase of consumer prices in the United States for each year since 1948, while the black line shows the annual rate of wage increase. The difference between the two indicates the rate of growth of *real* wages. Generally, wages rise faster than prices, reflecting the steady advance of technology and of labor productivity; therefore, real wages rise. (But this is not always the case; the graph shows that real wages fell several times in the 1980s.)

The feature of Figure 23-3 that virtually jumps off the page is the way the two lines dance together. Wages normally rise rapidly when prices rise rapidly, and

FIGURE 23-3 RATES OF CHANGE OF WAGES AND PRICES IN THE UNITED STATES, 1948–1995

This chart compares the rate of price inflation (orange line) with the rate of growth of nominal wages (black line) in the postwar period. The patterns are clearly quite similar, with wages and prices normally accelerating or decelerating together. Notice that wage increases generally outstrip price increases; that is, *real* wages normally rise from year to year. (*Note*: Data on both wages and prices pertain to the private business sector.)

SOURCE: Bureau of Labor Statistics.

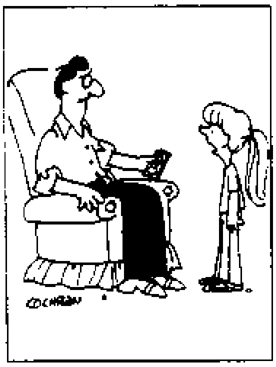


they rise slowly when prices rise slowly. But you should not draw any hasty conclusions from this association. It does not, for example, tell us whether rising prices cause rising wages or rising wages cause rising prices. Remember the warnings given in Chapter 1 about trying to infer causation just by looking at data. But analyzing cause and effect is not our purpose right now. We merely want to dispel the myth that inflation inevitably erodes real wages.

Why is this myth so widespread? Imagine a world without inflation in which wages are rising 2 percent a year because of the increasing productivity of labor. Now imagine that, all of a sudden, inflation sets in and prices start rising 3 percent a year but that nothing else changes. Figure 23-3 suggests that, with perhaps a small delay, wage increases will accelerate to 2 plus 3, or 5 percent a year.

Will workers view this change with equanimity? Probably not. To each worker, the 5 percent wage increase will be seen as something he earned by the sweat of his brow. In his view, he *deserves* every penny of his 5 percent raise. In a sense, he is right because "the sweat of his brow" earned him a 2 percent increment in real wages that, when the inflation rate is 3 percent, can only be achieved by increasing his money wages by 5 percent. An economist would divide the wage increase in the following way:

Reason for Wage Increase	Amount
Higher productivity	2%
Compensation for higher prices	3
Total	5%



"Sure, you're raising my allowance. But am I actually gaining any purchasing power?"

An item's **RELATIVE PRICE** is its price in terms of some other item rather than in terms of dollars.

But the worker will probably keep score differently. Feeling that he earned the entire 5 percent by his own merits, he will view inflation as having "robbed" him of three-fifths of his just desserts. The higher the rate of inflation, the more of his raise the worker will feel has been stolen from him.

Of course, nothing could be further from the truth. Basically, the economic system is rewarding the worker with *the same 2 percent real wage increment for higher productivity regardless of the rate of inflation*. The "evils of inflation" are often exaggerated because of a failure to understand this point.

A second reason for misunderstanding the effects of inflation is that people tend to think in terms of the number of dollars it takes to buy something rather than in terms of the *purchasing power* of these dollars. For example, if inflation doubles both prices and wages, workers will have to labor exactly the same amount of time as before to earn the price of a loaf of bread. But because they now pay $2 a loaf instead of $1, they feel that the price of bread is scandalously high. In fact, nothing really has changed; but people cling to an outmoded idea of what bread *should* cost.

A third misperception results from failure to distinguish between a *rise in the general price level* and a change in **relative prices**, that is, a rise in the price of one commodity relative to that of another. To see the distinction most clearly, imagine first a *pure inflation* in which *every* price rises by 10 percent during the year, so that relative prices do not change. Table 23-2 gives an example in which movie tickets go up from $6.00 to $6.60, candy bars from 50 cents to 55 cents, and automobiles from $9,000 to $9,900. After the inflation, just as before, it will still take 12 candy bars to buy a movie ticket, 1,500 movie tickets to buy a car, and so on. A person who manufactures candy bars in order to purchase movie tickets is neither helped nor harmed by the inflation. Neither is a car dealer with a sweet tooth.

But real inflations are not like this. When there is 10 percent general inflation—meaning that the "average price" rises by 10 percent—some prices may jump 20 percent or more while others actually fall.[3] Suppose that, instead of the price increases shown in Table 23-2, prices rise as shown in Table 23-3. Movie prices go up by 25 percent, but candy prices do not change. Surely, candy manufacturers who love movies will be disgruntled because it now costs 15 candy bars instead of 12 to get into the theater. They will blame inflation for raising the price of movie tickets, even though their real problem stems from the *increase in the price of movies relative to candy*. (They would have been hurt as much if movie tickets had remained at $6 while the price of candy fell to 40 cents.)

TABLE 23-2

Item	Last Year's Price	This Year's Price	Percentage Increase
Candy bar	$ 0.50	$ 0.55	10
Movie ticket	6.00	6.60	10
Automobile	9,000	9,900	10

[3]The way statisticians figure out "average" price increases is discussed in the appendix to this chapter.

TABLE 23-3

Item	Last Year's Price	This Year's Price	Percentage Increase
Candy bar	$ 0.50	$ 0.50	0
Movie ticket	6.00	7.50	25
Automobile	9,000	9,450	5

Since car prices have risen by only 5 percent, theater owners in need of new cars will be delighted by the fact that an auto now costs only 1,260 movie admissions—just as they would have cheered if car prices had fallen to $7,560 while movie tickets remained at $6.00. However, they are unlikely to attribute their good fortune to inflation—as indeed they should not. What has actually happened is that *cars became cheaper relative to movies.*

Because real-world inflations proceed at *uneven* rates, relative prices are constantly changing. There are gainers and losers, just as some would gain and others lose if relative prices were to change without any general inflation. Inflation, however, gets a bad name because losers often blame inflation for their misfortune while gainers rarely credit inflation for their good luck.

These three kinds of misconceptions may go a long way toward explaining why respondents to public opinion polls consistently list inflation as a major national issue, why higher inflation rates depress consumers, and why voters express their ire at the polls when inflation is high.

Inflation does not systematically erode the purchasing power of wages. Nor does it lead to "unfair" prices. Nor is it usually to blame when some goods become more expensive relative to others.

But not all of the costs of inflation are mythical. Let us now turn to some of the real costs.

INFLATION AS A REDISTRIBUTOR OF INCOME AND WEALTH

We have just seen that the *average* person is neither helped nor harmed by inflation. But almost no one is exactly average! Some people gain from inflation and others lose. It is hard to say anything more systematic than this about the effects of inflation on particular prices and wages.

But inflation does have systematic effects on the distribution of income and wealth. Senior citizens trying to scrape by on pensions or other fixed incomes suffer badly from inflation. Since they earn no wages, it is little solace to them that wages keep pace with prices. Their pension incomes do not.[4]

This example actually illustrates a much more general problem. Think of pensioners as people who "lend" money to an organization (the pension fund) when they are young in order to be "paid back" with interest when they are old.

[4]This is not, however, true of Social Security benefits, which are automatically increased to compensate recipients for changes in the price level.

Because of the rise in the price level during the intervening years, the unfortunate pensioners get paid back in less valuable dollars than those they originally loaned. In general:

Those who lend money are apt to be victimized by inflation.

While lenders may lose heavily, borrowers may do quite well. For example, homeowners who borrowed money from banks in the form of mortgages back in the 1950s, when interest rates were 3 or 4 percent, gained enormously from the surprisingly virulent inflation of the 1970s. They paid back dollars of much lower value than those that they borrowed. The same is true of other borrowers.

Borrowers often gain from inflation.

Since the redistribution caused by inflation generally benefits borrowers at the expense of lenders, and since both lenders and borrowers can be found at every income level, we conclude that:

Inflation does not always steal from the rich to aid the poor, nor does it always do the reverse.

Why, then, is the redistribution caused by inflation so widely condemned? Because its victims are selected capriciously. Nobody legislates the redistribution. Nobody enters into it voluntarily. The gainers do not earn their spoils, and the losers do not deserve their fate. Moreover, there are particular classes of people who inflation systematically robs of purchasing power year after year—people living on private pensions, families who save money and "lend" it to banks, and workers on long-term contracts or whose wages and salaries do not adjust easily for some other reason. Even if the average person suffers no damage from inflation, that offers little consolation to those who are hurt by it persistently and systematically. This is one fundamental indictment of inflation.

Inflation redistributes income in an arbitrary way. The actual income distribution should reflect the interplay of the operation of free markets and the purposeful efforts of government to alter that distribution. Inflation interferes with and distorts this process.

REAL VERSUS NOMINAL INTEREST RATES

But wait. Must inflation always rob lenders to bestow gifts upon borrowers? If both parties see inflation coming, won't lenders demand that borrowers pay a higher interest rate as compensation for the coming inflation? Indeed they will. For this reason, economists draw a sharp distinction between inflation that is *expected* and inflation that comes as a *surprise*.

What happens when inflation is fully expected by both parties? Suppose that Diamond Jim wants to borrow $1,000 from Scrooge, and both agree that, in the absence of inflation which erodes the purchasing power of money, a fair rate of interest would be 3 percent on a 1-year loan. This means that Diamond Jim would pay back $1,030 at the end of the year for the privilege of having $1,000 now.

If both expect prices to increase by 6 percent, Scrooge may reason as follows: "If Diamond Jim pays me back $1,030 a year from today, that money will buy less than what $1,000 buys today. Thus, I'll really be *paying him* to borrow from

me! I'm no philanthropist. Why don't I charge him 9 percent instead? Then he'll pay back $1,090 at the end of the year. With prices 6 percent higher, this will buy roughly what $1,030 is worth today. So I'll get the same 3 percent increase in purchasing power that we would have agreed on in the absence of inflation, and won't be any worse off. That's the least I'll accept."

Diamond Jim may follow a similar chain of logic. "With no inflation, I was willing to pay $1,030 a year from now for the privilege of having $1,000 today, and Scrooge was willing to lend it. He'd be crazy to do the same with 6 percent inflation. He'll want to charge me more. How much should I pay? If I offer him $1,090 a year from now, that will have roughly the same purchasing power as $1,030 today, so I won't be any worse off. That's the most I'll pay."

This kind of thinking may lead Scrooge and Diamond Jim to write a contract with a 9 percent interest rate—3 percent as the increase in purchasing power that Diamond Jim pays to Scrooge and 6 percent as compensation for expected inflation. Then, if the expected 6 percent inflation actually materializes, neither party will be better or worse off than was expected at the time the contract was signed.

The **REAL RATE OF INTEREST** is the percentage increase in purchasing power that the borrower pays to the lender for the privilege of borrowing. It indicates the increased ability to purchase goods and services that the lender earns.

The **NOMINAL RATE OF INTEREST** is the percentage by which the money the borrower pays back exceeds the money that he borrowed, making no adjustment for any fall in the purchasing power of this money that results from inflation.

This example illustrates a general principle. The 3 percent increase in purchasing power that Diamond Jim agrees to hand over to Scrooge is called the **real rate of interest**. The 9 percent contractual interest charge that Diamond Jim and Scrooge write into the loan agreement is called the **nominal rate of interest**. The nominal rate of interest is arrived at by adding the *expected rate of inflation* to the real rate of interest. Expected inflation is added to compensate the lender for the loss of purchasing power that he expects to suffer as a result of inflation. Because of this:

Inflation that is accurately predicted need not redistribute income between borrowers and lenders. If the *expected* rate of inflation that is embodied in the nominal interest rate closely approximates the *actual* rate of inflation, no one gains and no one loses. However, to the extent that expectations prove incorrect, inflation will still redistribute income.[5]

It need hardly be pointed out that errors in predicting the rate of inflation are the norm, not the exception. Published forecasts bear witness to the fact that economists have great difficulty in predicting the rate of inflation. The task is no easier for businesses, consumers, and banks. This is one reason why inflation is so widely condemned as unfair and undesirable. It sets up a guessing game that no one likes.

INFLATION AND THE TAX SYSTEM[6]

Inflation imposes costs on society because it is hard to predict. But other costs arise from inflation even when it is predicted accurately. These costs stem from the fact that many laws and regulations were designed for an inflation-free economy and may malfunction when inflation is high.

The tax system is probably the most important example. The law does not recognize the distinction between *nominal* and *real* interest rates; it simply taxes

[5]EXERCISE: Who gains and who loses if the inflation turns out to be only 4 percent instead of the 6 percent that Scrooge and Diamond Jim expected? What if the inflation rate is 8 percent?

[6]This section contains somewhat more difficult material that may be omitted from shorter courses.

TABLE 23-4 INFLATION AND THE TAXATION OF INTEREST INCOME

(1) Inflation Rate (percent)	(2) Nominal Interest Rate (percent)	(3) Interest Income (dollars)	(4) Loss of Purchasing Power Due to Inflation (dollars)	(5) Real Interest Income (dollars)	(6) Taxes Paid (dollars)	(7) Real Income After Tax (dollars)	(8) Real Income After Tax (as percentage of $1,000 loan)	(9) Effective Rate of Taxation (percent)
0	3	30	0	30	10	20	2	33⅓
6	9	90	60	30	30	0	0	100

nominal interest regardless of how much real interest it represents. As a result, strange things happen when there is high inflation. Our example of Scrooge's loan to Diamond Jim will illustrate the problem.

The top line of Table 23-4 shows how taxation affects the loan agreement when there is no inflation and the nominal and real interest rates are both 3 percent. Scrooge earns $30 in nominal interest income (Column 3). Since there is no inflation, this also represents $30 in real interest income (Column 5). If Scrooge pays one-third of his income in taxes, his tax bill rises by $10 (Column 6), leaving him with $20 after tax (Column 7). This $20 amounts to 2 percent of the $1,000 originally loaned (Column 8). Because his $10 tax payment is one-third of his $30 in real interest income, Scrooge's effective tax rate is 33-1/3 percent (Column 9), just as Congress intended.

Now let's consider the same transaction when the inflation rate is 6 percent and Scrooge and Diamond Jim settle on a 9 percent nominal interest rate. Scrooge collects $90 in interest (Column 3). But, with 6 percent inflation, the purchasing power of the $1,000 he lends declines by $60 (Column 4). Thus, his real interest income is again $30 (Column 5). However, the tax collector taxes the $90 *nominal* interest income, not the $30 *real* interest income, so Scrooge must pay $30 (one-third of $90) in taxes (Column 6). As we can see in Column 7, his after-tax real income on the loan is zero since the tax collector takes all of his real income. Thus, the effective tax rate on Scrooge's real interest income is 100 percent (Column 9), far larger than the 33-1/3 percent rate intended by Congress. In effect, Scrooge is being taxed on "phantom profits" that, in purchasing-power terms, he never received.

So a tax system that works well at zero inflation misfires at 6 percent inflation because it taxes nominal, rather than real, interest. This little example illustrates a pervasive and serious problem:

Because it fails to recognize the distinction between nominal and real interest rates, our tax system levies high, and presumably unintended, tax rates on interest income when there is high inflation. Similar problems arise in the taxation of capital gains, corporate profits, and other items. Some economists feel that these high tax rates discourage saving, lending, and investing and that high inflation therefore retards economic growth.

A **CAPITAL GAIN** is the difference between the price at which an asset is sold and the price at which it was bought.

A particularly acute version of this problem arises in the taxation of **capital gains**—the difference between the price at which an investor sells an asset and the price that she paid for it. An example will bring out the point. Between 1979

and 1996 the price level doubled, approximately. Consider some stock that was purchased for $5,000 in 1979 and sold for $7,500 in 1996. The investor actually *lost* purchasing power in the transaction because $7,500 in 1996 purchased less than $5,000 in 1979. Yet, since the law levies taxes on nominal capital gains, with no correction for inflation, the investor will be taxed on the $2,500 nominal capital gain as though there had been a profit rather than a loss.

Many economists have proposed that this, presumably unintended, feature of the law be changed by taxing only capital gains that exceed inflation. But, up to now, Congress has not agreed.

INTEREST RATE CEILINGS AND OTHER IMPEDIMENTS

These two quirks of the tax law illustrate a more general problem:

Many of the laws that govern our financial system become extremely counterproductive in an inflationary environment, causing problems that were never intended by the legislators.

Here are two more examples of laws that malfunction under inflation:

USURY LAWS Usury laws, which set *maximum* permissible interest rates on particular types of loans, date back to biblical days. But they can have strange and unintended consequences under inflation because they place ceilings on *nominal* interest rates rather than on *real* interest rates. In effect, limits on nominal interest rates make some activities which are perfectly legal at low or zero inflation illegal in inflationary environments. In fact, usury laws created so much havoc during the period of double-digit inflation in 1979–1980 that Congress took drastic action to curtail them.

REGULATION OF PUBLIC UTILITIES When the rate of inflation rose above 12 percent in 1980, there was a public uproar after regulated utilities asked the regulatory agencies to permit them to earn a rate of return closer to 11 percent—a *negative* real rate of return! They frequently found that their requests were considered exorbitant by the commissions and by the general public. The consequence was that many utilities could not afford to borrow the money needed to serve expanding public demand.

Thus, failure to understand that high *nominal* interest rates can signify low *real* interest rates has been known to make the tax code misfire, to impoverish savers, to inhibit borrowing and lending, and to make it nearly impossible for public utilities to raise the capital they need to serve rising consumer demands—with power shortages the predictable results. It is important to note that *these costs of inflation are not purely redistributive*. Society as a whole loses when mutually beneficial transactions are prohibited by obsolete legislation.

Why, then, do such laws stay on the books? One reason is a general lack of understanding of the difference between real and nominal interest rates. People fail to understand that it is normally the *real* rate of interest that matters in an economic transaction because only that rate reveals how much borrowers pay and lenders receive *in terms of the goods and services that money can buy*. They focus on the high *nominal* interest rates caused by inflation, even when these rates correspond to low real interest rates.

THE ILLUSION OF HIGH INTEREST RATES

The difference between real and nominal interest (and profit) rates, and the fact that the real rate matters economically while the nominal rate is politically significant, are matters that are of the utmost importance and yet are understood by very few people, including many who make public policy decisions in these areas.

This concept is one of our **Ideas for Beyond the Final Exam,** and if you remember it 10 years from now, you will truly have gotten a great deal out of studying economics.

OTHER COSTS OF INFLATION

Another cost of inflation is that rapidly changing prices make it risky to enter into long-term contracts. In an extremely severe inflation, the "long term" may be only a few days. But even moderate inflations can have remarkable effects on long-term loans. Suppose that a corporation wants to borrow $1 million to finance the purchase of some new equipment and needs the loan for 20 years. If inflation averages 4 percent over this period, the $1 million it repays at the end of 20 years will be worth $456,387 in today's purchasing power. If inflation averages 8 percent instead, it will be worth only $214,548. Lending or borrowing for this long a period is obviously a big gamble. With the stakes this high, the outcome may be that neither lenders nor borrowers want to get involved in long-term contracts. But without long-term loans, business investment often becomes impossible. The economy stagnates.

Inflation also makes life difficult for the shopper. You probably have a group of stores that you habitually patronize because they carry the items you want to buy at (roughly) the prices you want to pay. This knowledge saves you a great deal of time and energy. But when prices are changing rapidly, your list quickly becomes obsolete. You return to your favorite clothing store to find that the price of jeans has risen drastically. Should you buy? Should you shop around at other stores? Will they have also raised their prices? Business firms have precisely the same problem with their suppliers. Rising prices force them to shop around more, which imposes costs on the firms and, more generally, reduces the efficiency of the whole economy.

Shopping costs may sound frivolous and unimportant, but they are not. Arthur Okun, who chaired the Council of Economic Advisers under President Johnson, suggested an ingenious mental exercise that illustrates the importance of shopping costs. Ask yourself how much you would have to be paid to promise never again to buy anything from any of the stores you have patronized in the past. When you ponder this for a while, you realize the great value of having normal places to shop. Inflation takes some of this value away.

THE COSTS OF LOW VERSUS HIGH INFLATION

The preceding litany of the costs of inflation alerts us to one very important fact: *predictable inflation is far less burdensome than unpredictable inflation.* When is inflation most predictable? When it proceeds year after year at a modest and more or less steady rate. Thus, the *variability of the inflation rate* is a crucial factor. Inflation of 3 percent a year for three consecutive years will exact lower social costs

than inflation that is 2 percent in the first year, zero in the second, and 7 percent in the third. In general:

Steady inflation is more predictable than variable inflation and therefore has smaller social and economic costs.

But the *average level of the inflation rate* is also important. Partly because of the interest rate illusions mentioned above and partly because of the more rapid breakdown in normal customer relationships that we have just mentioned, a steady inflation of 6 percent a year does more damage than a steady inflation of 3 percent a year.

Economists distinguish between *low inflation*, which is a modest economic problem, and *high inflation*, which can be a devastating one, partly on the basis of the average level of inflation and partly on its variability. If inflation remains steady and low, prices may rise for a long time, but at a moderate and fairly constant pace, allowing people to adapt. From 1982 to 1995 in the United States, for example, prices climbed a total of 55 percent, for an average annual inflation rate of 3.4 percent. And the pace of inflation was remarkably steady, never dropping below 2.1 percent or rising above 4.4 percent.

Very high inflations typically last for shorter periods of time and are often marked by highly variable inflation rates from month to month or year to year. In recent decades, for example, countries ranging from Argentina to Israel to Russia have experienced bouts of inflation exceeding 100 percent or even 1,000 percent per year. (See the box on Nicaragua on the next page.) Each of these episodes severely disrupted the country's economy.

The German hyperinflation after World War I is perhaps the most famous episode of runaway inflation. Between December 1922 and November 1923, when a hard-nosed reform finally broke the spiral, wholesale prices in Germany increased by almost 100 million percent! But even this experience was dwarfed by the great Hungarian inflation of 1945–1946, the greatest inflation of them all. For a period of 1 year, the *monthly* rate of inflation averaged about 20,000 percent. In the final month, the price level skyrocketed 42 quadrillion percent!

If you review the costs of inflation that have been discussed in this chapter, you will see why the distinction between low and high inflation is so fundamental. Many economists think we can live rather well in an environment of steady, low inflation. No one believes we can survive very well under extremely high inflation.

When inflation is steady and low, the rate at which prices rise is relatively easy to predict. It then can be taken into account in setting interest rates (as long as the law allows). Under high inflation, especially if prices are rising at ever-increasing or highly variable rates, this is extremely difficult, and perhaps impossible, to do. The potential redistributions become monumental and, as a result, lending and borrowing may cease entirely.

Any inflation makes it difficult to write long-term contracts. Under low, creeping inflation, the "long term" may be 20 years, or 10 years, or 5. But under high, galloping inflation, the "long term" may be measured in days or weeks. Restaurant prices may change daily. Air fares may go up while you are in the middle of your journey. When it is impossible to enter into contracts of any duration longer than a few days, economic activity becomes paralyzed. We conclude that:

The horrors of hyperinflation either are absent in low, steady inflations or are present in such muted forms that they can scarcely be considered horrors.

HYPERINFLATION AND THE PIGGY BANK

While mild inflations are barely noticeable in everyday life, hyperinflation makes all sorts of normal economic activities more difficult and transforms a society in strange and unexpected ways. This article, excerpted from the *New York Times*, illustrates some of the problems that hyperinflation created for Nicaraguans in 1989.

For generations, Nicaraguans have guarded their savings in piggy banks. . . . But no longer. In a country where inflation recently reached 161 percent for a 2-week period, a penny saved is a penny spent. "No one wants a bank now," said a potter who has given over his kilns to making beer mugs. "We've given up even making them."

The demise of the piggy bank is only the least of the complications that have vexed the public as inflation and Government efforts to combat it have sent the value of the Nicaraguan córdoba fluctuating wildly.

After inflation became unbearable, the Government replaced all of its currency in February 1988 at a new rate of 10 córdobas to a dollar. But by December, the new money had reached 4,500 to the dollar, and despite months of harsh austerity measures, it has now surged upward again, reaching 26,250 to the dollar.

That kind of uncertainty has left the banking system in shambles, despite savings accounts that offer up to 70 percent interest a month. . . . And it has left a legacy of quirks that now extends throughout the country's daily life. . . .

In many parts of the country, enterprising mechanics have converted the nation's once-precious stock of coins into something more valuable: metal washers to fit the nuts and bolts of rapidly deteriorating machinery. . . .

In Managua, it is still necessary to deposit a copper-colored 1-córdoba coin to make a pay phone call. But . . . not everybody even remembers what a 1-córdoba coin looks like, and fewer still actually own one. That is probably just as well for the phone system, because if anyone bothered to carry the coins, they could make about 26,250 phone calls for a dollar. . . .

Beating the exchange rates is particularly trying for restaurant owners who must have any price increases approved by the Government Institute of Tourism. Since that process consumes precious amounts of time, restaurant owners must aim high with their requests in the anticipation that new inflation will make their prices competitive, but still profitable, at some future date.

When exchange rates are changing, therefore, prices at a given restaurant can shoot far out of sight, emptying it of patrons for days or weeks. Then, as has happened recently, a new devaluation can make the same prices absurdly low in dollar terms. A steak dinner for two at one of Managua's leading restaurants on a recent weekend cost a little over $3, if you could get a table.

SOURCE: Mark A. Uhlig, "Is Nicaraguan Piggy Bank an Endangered Species?" *New York Times*, June 24, 1989.

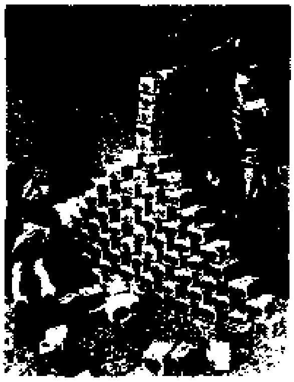


LOW INFLATION DOES NOT NECESSARILY LEAD TO HIGH INFLATION

We noted earlier that inflation is surrounded by a mythology that bears precious little relation to reality. It seems appropriate to conclude this chapter by disposing of one particularly persistent myth: that low inflation is a slippery slope that invariably leads to high inflation.

There is neither statistical evidence nor theoretical support for the myth that low inflation inevitably leads to high inflation. To be sure, inflations sometimes accelerate. But at other times they slow down.

While creeping inflations have many causes, runaway inflations have occurred only when the government has printed incredible amounts of money, usually to finance wartime expenditures. In the German inflation of 1923, the government



These children in Germany during the hyperinflation of the 1920s are building a pyramid with cash, worth no more than the sand or sticks used by children elsewhere.

finally found that its printing presses could not produce enough paper money to keep pace with the exploding prices. Not that it did not try. By the end of the inflation, the *daily* output of currency was over 400 quadrillion marks! The Hungarian authorities in 1945 to 1946 tried even harder. The average growth rate of the money supply was more than 12,000 percent *per month*. Needless to say, these are not the kind of inflation problems that are likely to face the United States in the foreseeable future.

But this should not be interpreted to imply that there is nothing wrong with low inflation. Much of this chapter has been spent analyzing the very real costs of any inflation, no matter how low. A case against even moderate inflation can indeed be built, but it does not help this case to shout foolish slogans like "Creeping inflation always leads to galloping inflation." Fortunately, it is simply not true.

SUMMARY

1. Unemployment exacts heavy financial and psychological costs from those who are its victims, costs that are borne quite unevenly by different groups in the population.
2. In recent decades, the U.S. economy often has produced less output than it could have produced were it operating at full employment, that is, at **potential GDP.**
3. **Frictional unemployment** arises when people are between jobs for normal reasons. Thus, most frictional unemployment is desirable.
4. **Structural unemployment** is due to shifts in the pattern of demand or to technological change that makes certain skills obsolete.
5. **Cyclical unemployment** is the portion of unemployment that rises in recessions and falls when the economy booms.
6. President Kennedy first enunciated the goal of 4 percent unemployment in 1961. But few economists think this is a realistic target for the 1990s. Most now think that *full employment* comes at an unemployment rate between 5.5 and 6.0 percent.
7. *Unemployment insurance* replaces nearly one-half of the lost income of unemployed persons who are insured. But only about one-third of the unemployed collect benefits, and no insurance program can bring back the lost output that could have been produced had these people been working.
8. People have many misconceptions about inflation. For example, many people believe that inflation systematically erodes **real wages,** are appalled by rising prices even when wages are rising just as fast, and blame inflation for any unfavorable changes in **relative prices.** All of these are myths.
9. Other costs of inflation are real, however. For example, inflation often redistributes income from lenders to borrowers.
10. This redistribution can be ameliorated by adding the *expected rate of inflation* to the interest rate. But expectations often prove to be inaccurate.
11. The **real rate of interest** is the **nominal rate of interest** minus the expected rate of inflation.
12. Since the real rate of interest indicates the command over real resources that the borrower surrenders to the lender, it is of primary economic importance.
13. Yet public attention often is riveted on nominal rates of interest, and this confusion can lead to costly policy mistakes when inflation converts high nominal interest rates into low real interest rates. This is one of the **Ideas for Beyond the Final Exam.**
14. Because nominal, not real, interest is taxed, our tax system levies heavy taxes on interest income when inflation is high.
15. Low inflation which proceeds at moderate and fairly predictable rates year after year carries far lower social costs than high and/or variable inflation. But even low, steady inflations entail costs.
16. The notion that low inflation inevitably accelerates into high inflation is a myth with no foundation in economic theory and no basis in historical fact.

KEY TERMS

Unemployment rate
Potential GDP
Labor force
Discouraged worker
Frictional unemployment
Structural unemployment
Cyclical unemployment
Full employment
Unemployment insurance
Purchasing power
Real wage rate
Relative price
Redistribution by inflation
Real rate of interest
Nominal rate of interest
Capital gain
Expected rate of inflation
Inflation and the tax system
Low versus high inflation

QUESTIONS FOR REVIEW

1. Why may it not be as terrible to become unemployed nowadays as it was during the Great Depression?
2. "Unemployment is no longer a social problem because unemployed workers receive unemployment benefits and other benefits that make up for most of their lost wages." Comment.
3. Using what you learned about aggregate demand and aggregate supply in the last chapter, try to explain why the U.S. economy has failed so frequently to produce up to its potential. (You will learn more about this question in later chapters, so don't worry if you find the question difficult now.)
4. Why is it so difficult to define *full employment*? What unemployment rate should the government be shooting for today?
5. Show why each of the following complaints is based on a misunderstanding about inflation:
   a. "Inflation must be stopped because it robs workers of their purchasing power."
   b. "Inflation is a terrible social disease. It leads to unconscionably high prices for basic necessities."
   c. "Inflation makes it impossible for working people to afford many of the things they were hoping to buy."
   d. "Inflation must be stopped today, for if we do not stop it, it will surely accelerate to ruinously high rates and lead to disaster."
6. What is the *real interest rate* paid on a credit-card loan bearing 14 percent nominal interest per year, if the rate of inflation is
   a. zero?
   b. 2 percent?
   c. 6 percent?
   d. 14 percent?
   e. 18 percent?
7. Suppose that you agree to lend money to your friend on the day you both enter college, at what you both expect to be a zero *real* rate of interest. Payment is to be made at graduation, with interest at a fixed *nominal* rate. If inflation proves to be *lower* during your 4 years in college than what you both had expected, who will gain and who will lose?
8. You have lived almost your entire life under moderate inflation. Can you think of some costs that inflation has imposed on you personally? How do these costs relate to the material in this chapter?
9. Add a third line to Table 23-4 showing what would happen if the inflation rate went to 12 percent and the real interest rate remained at 3 percent.

APPENDIX — HOW STATISTICIANS MEASURE INFLATION

INDEX NUMBERS FOR INFLATION

Inflation is generally measured by the change in some index of the general price level. For example, between 1973 and 1995, the Consumer Price Index (CPI) rose from 44.4 to 152.4, an increase of 243 percent. The meaning of the *change* is clear enough. But what are the meanings of the 44.4 figure for 1973 and the 152.4 figure for 1995? These numbers are *index numbers*.

An index number expresses the cost of a market basket of goods *relative to its cost in some "base" period.*

Since the CPI currently uses 1982–1984 as its base period, the CPI of 152.4 for 1995 means that it costs $152.40 to purchase the same basket of goods and services that cost $100 in 1982 to 1984.

Now, the particular basket of consumer goods and services under scrutiny really did not cost $100 in

1982 to 1984. When constructing index numbers, it is conventional to set the index at 100 in the base period. How is this conventional figure used in obtaining index numbers of other years? Very simply. Suppose that the budget needed to buy the roughly 250 items included in the CPI was $2,000 per month in 1982 to 1984 and $3,048 per month in 1995. Then the index is defined by the following rule:

$$\frac{\text{CPI in 1995}}{\text{CPI in 1982–1984}} = \frac{\text{Cost of the market basket in 1995}}{\text{Cost of the market basket in 1982–1984}}$$

Since the CPI in 1982–1984 is set at 100:

$$\frac{\text{CPI in 1995}}{100} = \frac{\$3{,}048}{\$2{,}000} = 1.524$$

or

$$\text{CPI in 1995} = 152.4$$

Exactly the same sort of equation enables us to calculate the CPI in any other year. We have the rule:

$$\text{CPI in given year} = \frac{\text{Cost of market basket in given year}}{\text{Cost of market basket in base year}} \times 100$$

Of course, not every combination of consumer goods that cost $2,000 in 1982 to 1984 rose to $3,048 by 1995. For example, a color TV set that cost $400 in 1983 might still have cost $400 in 1995, but a $400 hospital bill in 1983 might have ballooned to $1,200.

Since no two families buy precisely the same bundle of goods and services, no two families suffer precisely the same increase in their cost of living unless all prices rise at the same rate. Economists refer to this phenomenon as the **index number problem.**

When relative prices are changing, there is no such thing as a "perfect price index" that is correct for every consumer. Any statistical index will understate the increase in the cost of living for some families and overstate it for others. At best, the index can represent the situation of an "average" family.

THE CONSUMER PRICE INDEX

The *Consumer Price Index (CPI)*, which is calculated and announced each month by the Bureau of Labor Statistics (BLS), is surely the most closely watched price index. When you read in the newspaper or see on television that the "cost of living rose by 0.3 percent last month," chances are the reporter is referring to the CPI.

TABLE 23-5

RESULTS OF STUDENT EXPENDITURE SURVEY, 1983

Item	Average Price	Average Quantity Purchased per Month	Average Expenditure per Month
Hamburger	$ 0.80	70	$ 56
Jeans	24.00	1	24
Movie ticket	5.00	4	20
		Total	$100

The **Consumer Price Index (CPI)** is measured by pricing the items on a list representative of a typical urban household budget.

To know what items to include and in what amounts, the BLS conducts an extensive survey of spending habits roughly once every decade. This means that the *same* bundle of goods and services is used as a standard for 10 years or so, whether or not spending habits change.[7] Of course, spending habits do change, and this introduces a small error into the CPI's measurement of inflation.

A simple example will help us understand how the CPI is constructed. Imagine that college students purchase only three items—hamburgers, jeans, and movie tickets—and that we want to devise a cost-of-living index (call it SPI, or "student price index") for them. First, we would conduct a survey of spending habits in the base year. (Suppose it is 1983.) Table 23-5 represents the hypothetical results. You will note that the frugal students of that day spent only $100 per month: $56 on hamburgers, $24 on jeans, and $20 on movies.

Table 23-6 presents hypothetical prices of these same three items in 1996. Each price has risen by a different amount, ranging from 25 percent for jeans up to 50 percent for hamburgers. By how much has the SPI risen? Pricing the 1983 student budget at 1996 prices, we find that what once cost $100 now

[7]Economists call this a *base-period weight index* because the relative importance it attaches to the price of each item depends on how much money consumers actually chose to spend on the item during the base period.

TABLE 23-6

PRICES IN 1996

Item	Price	Percentage Increase over 1983
Hamburger	$ 1.20	50
Jeans	30.00	25
Movie ticket	7.00	40

TABLE 23-7

COST OF 1983 STUDENT BUDGET IN 1996 PRICES

70 hamburgers at $1.20	$ 84
1 pair of jeans at $30	30
4 movie tickets at $7	28
Total	$142

costs $142, as the calculation in Table 23-7 shows. Thus, the SPI, based on 1983 = 100, is:

$$\text{SPI} = \frac{\text{Cost of budget in 1996}}{\text{Cost of budget in 1983}} \times 100$$

$$= \frac{\$142}{\$100} \times 100 = 142$$

So the SPI in 1996 stands at 142, meaning that students' cost of living has increased 42 percent over the 13 years.

HOW TO USE A PRICE INDEX TO "DEFLATE" MONETARY FIGURES

One of the most common uses of price indexes is in the comparison of monetary figures relating to two different points in time. The problem is that, if there has been inflation, the dollar is not a good measuring rod because it can buy less now than it did in the past.

Here is a simple example. Suppose that the average student spent $100 per month in 1983 but $130 per month in 1996. If there was an outcry that students had become spendthrifts, how would you answer the charge?

The obvious answer is that a dollar in 1996 does not buy what it did in 1983. Specifically, our SPI shows us that it takes $1.42 in 1996 to purchase what $1 would purchase in 1983. To compare the spending habits of students in the 2 years, we must divide the 1996 spending figure by 1.42. Specifically, *real* spending per student in 1996 (where "real" is defined by 1983 dollars) is:

$$\text{Real spending in 1996} = \frac{\text{Nominal spending in 1996}}{\text{Price index of 1996}}$$

Thus,

$$\text{Real spending in 1996} = \frac{\$130}{1.42} = \$91.55$$

This calculation shows that, despite appearances to the contrary, the change in nominal spending from $100 to $130 actually represented a *decrease* in real spending.

This calculation procedure is called *deflating by a price index*, and it serves to translate noncomparable monetary figures into more directly comparable real figures.

Deflating is the process of finding the real value of some monetary magnitude by dividing by some appropriate price index.

A good practical illustration is the real wage, a concept we have discussed in this chapter. Average hourly earnings in the U.S. economy were $8.02 in 1983 and $11.46 in 1995. Since the CPI in 1995 was 152.4 (with 1982–1984 as the base period), the real wage in 1995 (expressed in 1982–1984 dollars) was:

$$\text{Real wage in 1995} = \frac{\text{Money wage in 1995}}{\text{Price index of 1995}}$$

$$= \frac{\$11.46}{152.4} \times 100 = \$7.52$$

Thus, by this measure, the real wage fell 6.2 percent over the 12 years.

THE GDP DEFLATOR

In macroeconomics, one of the most important of the monetary magnitudes that we have to deflate is the nominal gross domestic product (GDP).

The price index used to deflate GDP is called the **GDP deflator.**

YOUR PERSONAL CPI

[illegible] impacts people differ[illegible] because everyone has some[illegible] unique spending habits. For [illegible], many senior citizens [illegible] heavily on medical care, but their homes have long been paid off. Many young professionals, on the other hand, have heavy mortgage payments but almost no doctor bills. Thus, in principle, everyone has his or her own personal CPI.

Use the accompanying chart to calculate yours for the year 1994. Just enter the percentage of your spending that you allocated to each category shown in the table. National inflation figures are based on the assumption that the typical household allocates 41 percent of its spending to housing. But if you spent only 30 percent on housing, enter 30 in the second column and multiply it by the inflation rate for housing listed in the third column. Enter the result in the far right column. Then add up the final column to obtain your personal inflation rate. The national inflation rate for 1994 was around 3 percent. How did you do?

Expenditure Category	Percentage You Spend per Category	Inflation Rate for This Item	Contribution to Your Inflation Rate
Food and beverages		2.3%	
Housing		2.5	
Apparel		0.7	
Transportation		3.3	
Medical care		4.7	
Entertainment		2.5	
Miscellaneous		4.4	
Total			

SOURCE: "What's Your Inflation Rate?" *Kiplinger's Personal Finance Magazine*, January 1995, p. 59.

Our general principle for deflating a nominal magnitude tells us how to go from nominal GDP to real GDP:

$$\text{Real GDP} = \frac{\text{Nominal GDP}}{\text{GDP deflator}} \times 100$$

As with the CPI, the 100 simply serves to establish the base of the index as 100, rather than 1.00.

Some economists consider the GDP deflator to be a better measure of overall inflation in the economy than the Consumer Price Index. The main reason for this is that the two price indexes are based on different market baskets. As already mentioned, the CPI is based on the budget of a typical urban family. By contrast, the GDP deflator is constructed from a market basket that includes *every* item in the GDP—that is, every final good and service produced by the economy. Thus, in addition to prices of consumer goods, the GDP deflator includes the prices of airplanes, lathes, and other goods purchased by businesses. It also includes government services. For this reason, the measures of inflation that these two indexes give are rarely the same. Usually their disagreements are minor. But sometimes they can be substantial, as in 1980 when the CPI recorded a 13.5 percent inflation rate over 1979 while the GDP deflator recorded only 9.2 percent.

SUMMARY

1. Inflation is measured by the percentage increase in an **index number** of prices, which shows how the cost of some basket of goods has changed over a period of time.
2. Since relative prices are changing all the time, and since different families purchase different items, no price index can represent precisely the change in the cost of living for every family.
3. The **Consumer Price Index (CPI)** tries to measure the cost of living for an "average" urban household by pricing a "typical" market basket every month.
4. Price indexes like the CPI can be used to **deflate** monetary figures to make them more comparable. This amounts to dividing the monetary magnitude by the appropriate price index.
5. The **GDP deflator** is a better measure of economy-wide inflation than the CPI because it includes the prices of all goods and services in the economy.

KEY TERMS

Index number
Index number problem
Consumer Price Index
Deflating
GDP deflator

QUESTIONS FOR REVIEW

1. Just below, you will find the yearly average values of the Dow Jones Industrial Average, the most popular index of stock market prices, for 4 different years. The Consumer Price Index for each year (on a base of 1982–1984 = 100) can be found on the inside back cover of this book. Use these numbers to deflate all four stock market values. In which year were stocks really worth the most?

Year	Dow Jones Industrial Average
1964	834
1972	951
1987	2,276
1994	3,794

2. Just below you will find nominal GDP and the GDP deflator for 1974, 1984, and 1994.
   a. Compute real GDP for each year.
   b. Compute the percentage change in nominal and real GDP from 1974 to 1984, and from 1984 to 1994.
   c. Compute the percentage change in the GDP deflator over these two periods.

GDP Statistics	1974	1984	1994
Nominal GDP (billions of dollars)	1,497	3,902	6,931
GDP Deflator	38.5	75.9	105.0

3. Fill in the blanks in the following table of GDP statistics:

	1992	1993	1994
Nominal GDP	6,244		6,931
Real GDP	6,244	6,384	
GDP Deflator		102.6	105.0

4. Use the following data to compute the College Price Index for 1996 using the base 1972 = 100.

Item	Price in 1972	Quantity per Month in 1972	Price in 1996
Button-down shirts	$10	1	$25
Loafers	25	1	55
Sneakers	10	3	35
Textbooks	12	12	40
Jeans	12	3	30
Restaurant meals	5	11	14

5. Average hourly earnings in the U.S. economy during several past years were as follows:

1964	1974	1984	1994
$2.36	$4.24	$8.32	$11.13

   Use the CPI numbers provided on the inside back cover to calculate the real wage (in 1982–1984 dollars) for each of these years. Which decade had the fastest growth of money wages? Which had the fastest growth of real wages?

6. The example in the appendix showed that the Student Price Index (SPI) rose by 42 percent from 1983 to 1996. You can understand the meaning of this better if you:
   a. Use Table 23-5 to compute the fraction of total spending accounted for by each of the three items in 1983. Call these the "expenditure weights."
   b. Compute the weighted average of the percentage increases of the three prices shown in Table 23-6, using the expenditure weights you have just computed.
   c. You should get 42 percent as your answer. This shows that "inflation," as measured by the SPI, is a weighted average of the percentage price increases of all the items that are included in the index.



CHAPTER 24

INCOME AND SPENDING: THE POWERFUL CONSUMER

Men are disposed, as a rule and on the average, to increase their consumption as their income increases, but not by as much as the increase in their income.
John Maynard Keynes

In Chapter 22, we noted that the strength of aggregate demand influences the performance of the economy. When aggregate demand is growing briskly, the economy is likely to be booming, though it may also be having trouble with inflation. When aggregate demand stagnates, a recession may follow.

This chapter begins our detailed study of the theory of income determination, the tool economists use to analyze issues like these. The theory is based on the concepts of aggregate demand and supply. In this and the next two chapters, we construct a simplified model of aggregate demand and learn where the *aggregate demand curve* of Chapter 22 comes from. Then Chapter 27 completes the model by adding the *aggregate supply curve*.

This first model of the macroeconomy can teach us much about the causes of unemployment and inflation. But it is too simple to deal with policy issues because the government and the financial system are largely ignored. We remedy these omissions in Part VII, where we give government spending, taxation, and interest rates appropriately prominent roles. The influence of the exchange rate between the U.S. dollar and foreign currencies is considered in Part VIII.

We build our model in steps. Since consumer spending accounts for the lion's share of total demand, it is natural to begin there. First, we need some definitions of alternative concepts of economic activity—distinguishing carefully among total *spending* (aggregate demand), total *output*, and total *income*. Next, we turn to the interactions among these three concepts, using a convenient pictorial device that shows how they are all related. The bulk of the chapter, however, is devoted to describing and analyzing the important relationship between consumer income and consumer spending. Finally, we discuss some complications that arise from the fact that consumer income, though crucial, is not the *only* factor governing consumer spending.

A PUZZLE: DEMAND MANAGEMENT AND THE ORNERY CONSUMER

We suggested in Chapter 22 that the government sometimes wants to shift the aggregate demand curve. There are a number of ways in which it can try to do so. One direct approach is to alter its own spending, becoming extravagant when private demand is weak and miserly when private demand is strong. But the government can also take a more indirect route by using taxes and other policy tools to influence *private* spending decisions. Since consumer expenditures constitute about two-thirds of gross domestic product, the consumer presents the most tempting target.

While there are many things it can do to alter consumer spending, the government's principal weapon is the personal income tax. Many of you already have encountered Form 1040, the unwelcome New Year's greeting that every taxpayer receives from the federal government each January. Many more of you probably have been on a payroll and seen a share of your wages deducted and sent to the Internal Revenue Service. It should be no mystery, then, how changes in personal taxes affect consumer spending. Any reduction in personal taxes leaves consumers with more after-tax income to spend. Any increase in taxes leaves less.

The linkage from taxes to spendable income to consumer spending seems direct and unmistakable, and, in a certain sense, it is. But a look at the history of some major tax changes aimed at altering consumer spending is sobering. The varying degrees of success both of the measures themselves and of the predictions of their effects explain why economic research into the relationship between taxes and consumption continues.

CASE 1: THE 1964 TAX CUT

The year 1964 was a good one for economists. For years they had been proclaiming that a cut in personal income taxes would be an excellent way to stimulate a stagnating economy. But the plea fell on deaf ears until President John F. Kennedy was persuaded of the basic logic of the argument and his successor, Lyndon Johnson, pushed the legislation through Congress. The 1964 tax cut was designed to spur consumer spending, and it succeeded admirably. Consumers reacted just about as the textbooks of the day predicted, the economy improved rapidly and markedly, and economists smiled knowingly.

CASE 2: THE 1975 TAX CUT

The next major attempt to stimulate the economy by cutting taxes met with much less success. In the spring of 1975, as the economy hit the bottom of a recession, President Gerald Ford and Congress agreed on a temporary tax cut to spur consumer spending: They refunded to each taxpayer part of the taxes paid in 1974, and reduced income tax rates for the rest of 1975. However, consumers confounded the wishes of the president and Congress by saving a good deal of their rebates rather than spending them.

CASE 3: THE 1981–1984 TAX CUTS

When President Ronald Reagan was elected in 1980, one of the first things he did was to push a series of reductions in personal income tax rates through Congress. Tax rates fell by about 23 percent between 1981 and 1984, and consumer

spending increased by more or less the amounts that economists predicted, thereby contributing to the long economic expansion of the 1980s.

AGGREGATE DEMAND is the total amount that all consumers, business firms, and government agencies are willing to spend on final goods and services.

Thus, tax policy did more or less what it was expected to do in 1964 and 1981 to 1984 but seemed to be less effective in 1975. Why? This chapter will attempt to provide some answers. But before getting involved in such complicated issues, we must build some vocabulary and learn some basic concepts. Then, at the end of the chapter, we will see what went wrong in 1975.

AGGREGATE DEMAND, DOMESTIC PRODUCT, AND NATIONAL INCOME

CONSUMER EXPENDITURE, symbolized by the letter ***C***, is the total amount spent by consumers on newly produced goods and services (excluding purchases of new homes, which are considered investment goods).

INVESTMENT SPENDING, symbolized by the letter ***I***, is the sum of the expenditures of business firms on new plant and equipment and households on new homes. Financial "investments" are not included, nor are resales of existing physical assets.

GOVERNMENT PURCHASES, symbolized by the letter ***G***, refer to the goods (such as airplanes and paper clips) and services (such as school teaching and police protection) purchased by all levels of government.

NET EXPORTS, symbolized by ***(X − IM)***, is the difference between U.S. exports and U.S. imports. It indicates the difference between what we sell to foreigners and what we buy from them.

First, the vocabulary. We have already introduced the concept of *gross domestic product* as the standard measure of the economy's total output.[1] For the most part, goods are produced in a market economy only if firms think they can sell them. **Aggregate demand,** another concept encountered in Chapter 22, is the total amount that all consumers, business firms, government agencies, and foreigners wish to spend on all U.S. final goods and services.

The downward-sloping aggregate demand curve of Chapter 22 alerted us to the fact that *aggregate demand is a schedule, not a fixed number.* The actual numerical value of aggregate demand depends on the price level, and several reasons for this dependence will emerge in coming chapters.

But the level of aggregate demand also depends on a variety of other factors like consumer incomes, various government policies, and events in foreign countries. We can understand the nature of aggregate demand best if we break it up into its major components.

Consumer expenditure (*consumption* for short) is simply the total demand for all consumer goods and services. This is the focus of the current chapter, and we represent it by the letter *C*. Consumer expenditure constitutes about two-thirds of total spending.

Investment spending, represented by the letter *I*, is the amount that firms spend on factories, machinery, and the like, plus the amount that families spend on new houses. Notice that this usage of the word *investment* differs from common parlance. Most people speak of "investing" in the stock market or in a bank account. This kind of "investment" merelv swaps one form of financial asset (such as money) for another form (such as a share of stock). When economists speak of "investment," they mean instead the purchase of some *new, physical* asset, like a drill press or an oil rig or a house. It is only these kinds of investments that lead directly to additional demand for newly produced goods and, subsequently, to greater productive capacity.

The third major component of aggregate demand is **government purchases** of goods and services—items like paper, computers, airplanes, ships, and labor that are bought by all levels of government. We will use the symbol *G* for this variable.

The final component of aggregate demand is **net exports,** which are simply defined as U.S. exports *minus* U.S. imports. The reasoning here is simple. Part of the demand for American goods and services originates beyond our borders—as when foreigners buy our wheat, our computers, and our banking services. So this must be added to domestic demand. Similarly, some items included in *C*

[1] See Chapter 22, pages 521–525.

and I are not American made. Think, for example, of German beer, Japanese cars, and Korean textiles. So these must be subtracted if we want to measure total spending on U.S. products. The addition of exports, which we represent by the symbol X, and the subtraction of imports, IM, leads us to the following shorthand definition of aggregate demand:

NATIONAL INCOME is the sum of the incomes that all individuals in the economy earned in the forms of wages, interest, rents, and profits. It excludes transfer payments and is calculated before any deductions are taken for income taxes.

Aggregate demand is the sum $C + I + G + (X - IM)$.

The last concept we need for our vocabulary is a way to measure the total *income* of all individuals in the economy. There are two versions of this: one for before-tax incomes, called **national income**, and one for after-tax incomes, called **disposable income.**[2] The term *disposable income*, which we will abbreviate DI, will play a prominent role in this chapter because it tells us how many dollars consumers actually have available to spend or to save.

THE CIRCULAR FLOW OF SPENDING, PRODUCTION, AND INCOME

DISPOSABLE INCOME is the sum of the incomes of all the individuals in the economy after all taxes have been deducted and all transfer payments have been added.

Enough definitions. How do these three concepts—domestic product, total expenditure, and national income—interact in a market economy? We can answer this best with a rather elaborate diagram (Figure 24-1). For obvious reasons, Figure 24-1 is called a *circular flow diagram*. It depicts a large, circular tube in which an imaginary fluid is circulating in a clockwise direction. There are several breaks in the tube where either some of the fluid leaks out or additional fluid is injected in.

Let us examine this system, beginning on the far left. At point 1 on the circle, we find consumers. Disposable income (DI) is flowing into them, and two things are flowing out: consumption (C), which stays in the circular flow, and saving (S), which "leaks out." This depicts the fact that consumers normally spend less than they earn and save the balance. The "leakage" to savings, of course, does not disappear, but flows into the financial system via banks, mutual funds, and so on. We postpone consideration of what happens there until Chapter 29.

The upper loop of the circular flow represents expenditures, and as we move clockwise to point 2, we encounter the first "injection" into the flow: investment spending (I). The diagram shows this as coming from "investors"—a group that includes both business firms and consumers.[3] As the circular flow moves past point 2, it is bigger than it was before: Total spending has increased from C to $C + I$.

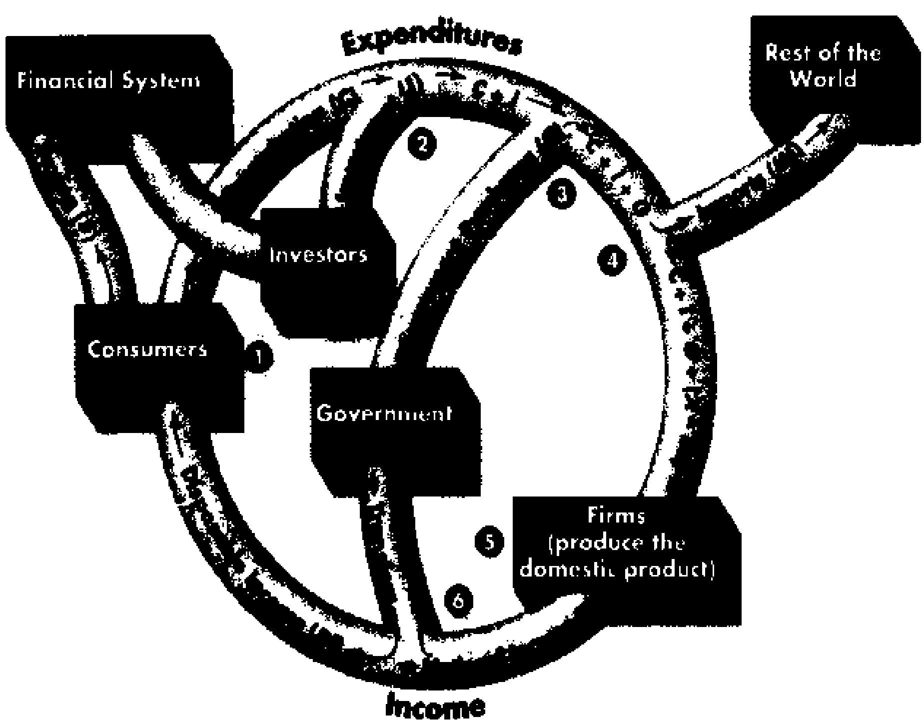
At point 3, there is yet another injection. The government adds its demand for goods and services (G) to those of consumers and investors ($C + I$). Now aggregate demand is up to $C + I + G$.

The final leakage and injection come at point 4. Here we see export spending entering the circular flow from abroad and import spending leaking out. The net effect of these two forces may increase or decrease the circular flow, depending on whether net exports are positive or negative. In either case, by the time we pass point 4, we have accumulated the full amount of aggregate demand, $C + I + G + (X - IM)$.

[2]More detailed information on these and other concepts is provided in the appendix to this chapter.

[3]You are reminded that expenditure on housing is part of I, not part of C.

FIGURE 24-1 THE CIRCULAR FLOW OF EXPENDITURES AND INCOME



The upper part of this circular flow diagram depicts the flow of expenditures on goods and services that comes from consumers (point 1), investors (point 2), government (point 3), and foreigners (point 4); that flow goes to the firms that produce the output (point 5). The lower part of the diagram indicates how the income paid out by firms (point 5) flows to consumers (point 1), after some is siphoned off by the government in the form of taxes and part of this is replaced by transfer payments (point 6).

The circular flow diagram shows this aggregate demand for goods and services arriving at the business firms, which are located at point 5. Responding to this demand, firms produce the domestic product. As the circular flow emerges from the firms, however, we have renamed it *national income*. Why? The reason is that, except for some complications explained in the appendix:

National income and domestic product must be equal.

Why is this so? When a firm produces and sells $100 worth of output, it pays most of the proceeds to its workers, to people who have lent it money, and to the landlord who owns the property on which it is located. All of these payments are *income* to some individuals. But what about the rest? Suppose, for example, that the firm pays wages, interest, and rent totaling $90 million and it sells its output for $100 million. What happens to the remaining $10 million? The answer is that the owners of the firm receive it as *profits*. Since these owners are citizens of the country, their incomes also count in national income.[4] Thus, when we add up all the wages, interest, rents, *and profits* in the economy to obtain the national *income*, we must arrive at the *value of output*.

TRANSFER PAYMENTS are sums of money that certain individuals receive as outright *grants* from the government rather than as payments for services rendered to employers. Some common examples are Social Security and unemployment benefits.

The lower loop of the circular flow diagram traces the flow of income by showing national income leaving the firms and heading for consumers. But there is a detour along the way. At point 6, the government does two things. First, it siphons off a portion of the national income in the form of taxes. Second, it adds back government **transfer payments**, like unemployment compensation and

[4]Some of the income paid out by American companies goes to non-citizens. Similarly, some Americans earn income from foreign firms. This complication is dealt with in the chapter appendix.

Social Security benefits, which are funds that certain individuals receive as outright *grants* from the government rather than as payments for goods or services rendered.

By subtracting taxes from GDP and adding transfer payments, we obtain disposable income.[5]

$$DI = \text{GDP} - \text{Taxes} + \text{Transfer payments}$$
$$= \text{GDP} - (\text{Taxes} - \text{Transfers})$$

or

$$DI = Y - T$$

where Y represents GDP and T represents taxes *net of transfers*. Disposable income flows unimpeded to consumers at point 1, and the cycle repeats.

Figure 24-1 raises several complicated questions which we pose now but will not try to answer at this early stage. The answers will be made clear in subsequent chapters.

1. Is the flow of spending and income growing larger or smaller as we move clockwise around the circle? Why?
2. Is the output that the firms produce at point 5 (the GDP) equal to aggregate demand? If so, what makes these two quantities equal? If not, what happens?

The next chapter provides the answers to these two questions.

3. Are the government's accounts in balance, so that what flows in at point 6 (taxes minus transfers) is equal to what flows out at point 3 (government purchases)? What happens if they are not?

This important question is first addressed in Chapter 28 and then recurs many times, especially in Chapter 32, which is devoted to discussing budget deficits.

4. Is our international trade balanced, so that exports equal imports? More generally, what factors determine net exports and what are the consequences of trade deficits or surpluses?

We take up these questions briefly in the next two chapters and then consider them fully in Part VIII, which is devoted to international economic issues.

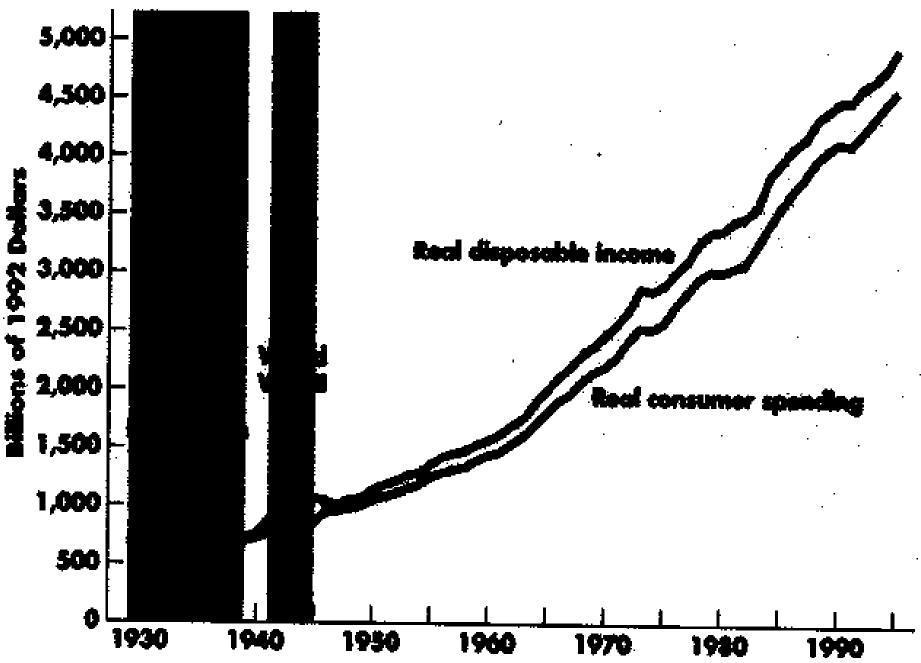
However, we cannot discuss any of these issues profitably until we first understand what goes on at point 1, where consumers make decisions. We turn next, therefore, to the determinants of consumer spending.

CONSUMER SPENDING AND INCOME: THE IMPORTANT RELATIONSHIP

Recall that we started the chapter with a puzzle: Why did consumers respond more or less as expected to tax changes in 1964 and the early 1980s, but not in 1975? An economist interested in predicting how consumer spending will respond to a change in personal income tax payments must first ask how C is related to disposable income; for an increase in taxes is a decrease in after-tax income, and a reduction in taxes is an increase in after-tax income. This section, therefore, will examine what we know about the response of consumer spending to a change in disposable income.

[5]This definition omits a few minor details, which are explained in the appendix to this chapter.

FIGURE 24-2 CONSUMER SPENDING AND DISPOSABLE INCOME



This time-series chart shows the behavior of consumer spending and disposable income in the United States since 1929. Except for the World War II years, the correspondence between the two variables is remarkably close. The distance between the two lines represents consumer saving, which was obviously quite small during the Great Depression of the 1930s and quite large during World War II.

SOURCE: U.S. Department of Commerce.

Figure 24-2 depicts the historical paths of C and *DI* for the United States since 1929. The association is obviously rather close and certainly suggests that consumption will rise whenever disposable income does, and fall whenever income falls. The difference between the two lines is personal saving. Notice how little saving consumers did during the Great Depression of the 1930s, where the two lines are very close together, and how much they did during World War II, when many consumer goods were either unavailable or rationed, leaving little on which to spend money.

Of course, knowing that consumer expenditures, C, will move in the same direction as disposable income, *DI*, is not enough for policy planners. They need to know *how much* one will go up when the other rises a given amount. Figure 24-3 presents the same data as in Figure 24-2, but in a way designed to help answer the "how much" question.

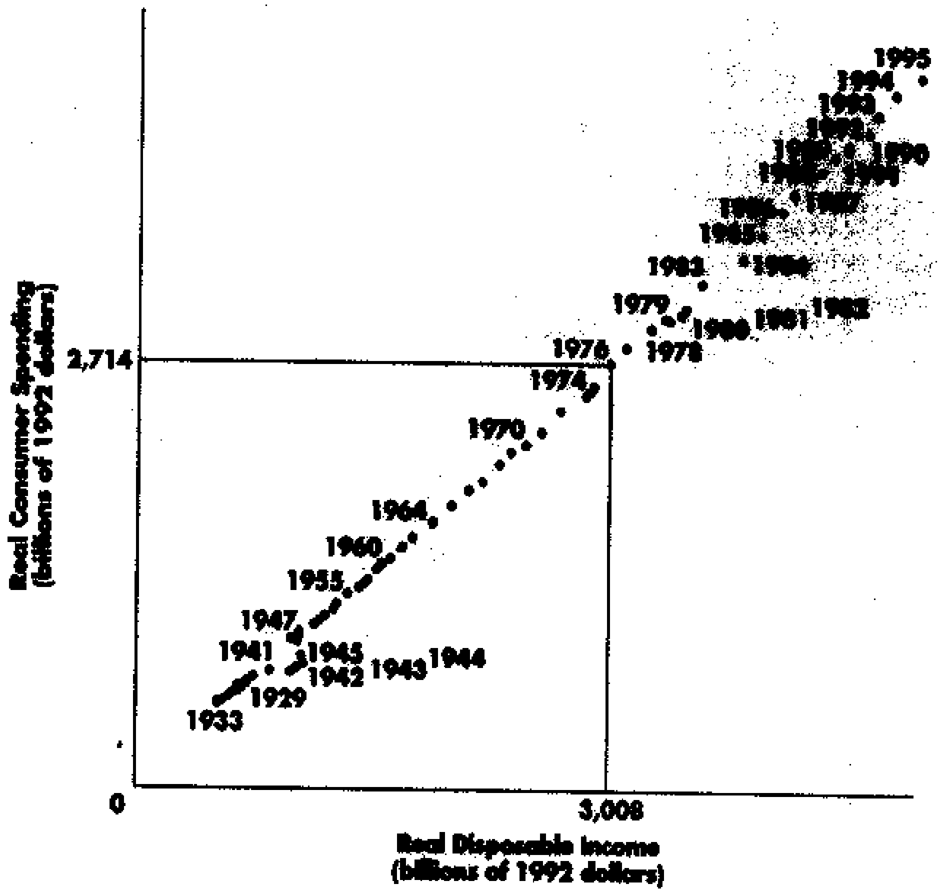
A **SCATTER DIAGRAM** is a graph showing the relationship between two variables (such as consumer spending and disposable income). Each year is represented by a point in the diagram. The coordinates of each year's point show the values of the two variables in that year.

Economists call such pictures **scatter diagrams**, and these diagrams are very useful in predicting how one economic variable (in this case, consumer spending) will change in response to a change in another economic variable (in this case, disposable income). Each dot in the diagram represents the data on C and *DI* corresponding to a particular year. For example, the point labeled "1976" shows that real consumer expenditures in 1976 were $2,714 billion (which we read off the vertical axis), while real disposable incomes amounted to $3,008 billion (which we read off the horizontal axis). Similarly, each year from 1929 to 1995 is represented by its own dot in Figure 24-3.

How can such a diagram assist the fiscal policy planner? Imagine that this is 1963 and you must decide whether to recommend to Congress a tax cut of $5 billion, $10 billion, or $15 billion. You have forecasts of what consumer expenditures are expected to be if taxes are not reduced. This, plus other forecasts of investment, government spending, and net exports has led you to conclude that aggregate demand in 1964 will be insufficient if taxes are not reduced.

FIGURE 24-3 SCATTER DIAGRAM OF CONSUMER SPENDING AND DISPOSABLE INCOME

This diagram shows the same data as depicted in Figure 24-2 but in a different manner. Each point on the diagram represents the data for both consumer spending and disposable income during a particular year. For example, the point labeled "1976" indicates that in that year consumer spending was $2,714 billion while disposable income was $3,008 billion. Diagrams like this one are called *scatter diagrams*.



To assist your imagination, another scatter diagram is given in Figure 24-4. This one removes the points for 1964 through 1995, which appear in Figure 24-3; after all, these were not known in 1963. Years prior to 1947 have also been removed because, as Figure 24-2 showed, both the Great Depression and wartime rationing seriously disturbed the normal relationship between *DI* and *C* during that period. With the training in economics that you have right now, what would you do?

One rough-and-ready approach is to get a ruler, set it down on Figure 24-4, and sketch a straight line that comes as close as possible to hitting all the points. That has been done for you in the figure, and you can see that it comes remarkably close to touching all of the points. The line summarizes, in a very rough way, the consumption-income relationship that is the focus of this chapter. The two variables do indeed appear to be closely related.

The *slope* of the straight line in Figure 24-4 is very important.[6] Specifically, we note that it is:

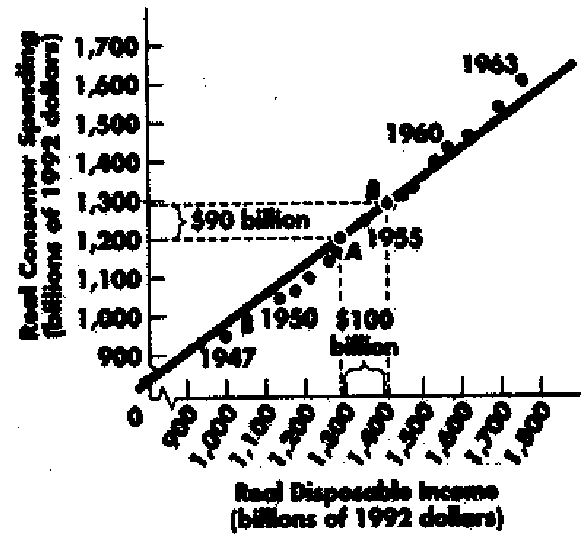
$$\text{Slope} = \frac{\text{Vertical change}}{\text{Horizontal change}} = \frac{\$90 \text{ billion}}{\$100 \text{ billion}} = 0.90$$

Since the horizontal change involved in the move from *A* to *B* represents a rise in disposable income of $100 billion (from $1,300 billion to $1,400 billion), and the corresponding vertical change represents the associated $90 billion rise in

[6] To review the concept of *slope*, turn back to page 19.

FIGURE 24-4 SCATTER DIAGRAM OF CONSUMER SPENDING AND DISPOSABLE INCOME, 1947–1963

This scatter diagram omits some of the data found in Figure 24-3 and thereby indicates the information that policy planners might have used in deciding upon the size of the 1964 income tax cut. The straight line in orange comes about as close as possible to fitting all the data points.



consumer spending (from $1,200 billion to $1,290 billion), the slope of the line indicates how spending responds to changes in disposable income. In this case, we see that each additional $1 of income leads to 90 cents of additional spending.

In terms of the policy issue of 1964, this line can therefore help provide an answer to the question: How much more consumer spending will be induced by tax cuts of $5 billion, $10 billion, or $15 billion, if the effects are similar to those observed in the past? First, we need to keep in mind that each dollar of tax cut increases disposable income by $1. Then we apply the finding from Figure 24-4 that each additional dollar of disposable income increases consumer spending by 90 cents to conclude that proposed tax cuts of $5 billion, $10 billion, or $15 billion would be expected to increase consumer spending by $4.5 billion, $9.0 billion, and $13.5 billion, respectively. Similar questions addressed by economists in 1964 led to a decision to cut taxes by about $9 billion.

THE CONSUMPTION FUNCTION AND THE MARGINAL PROPENSITY TO CONSUME

The **CONSUMPTION FUNCTION** is the relationship between total consumer expenditures and total disposable income in the economy, holding all other determinants of consumer spending constant.

It has been said that economics is just systematized common sense. Let us, then, try to organize and generalize what has been a completely intuitive discussion thus far. One thing we have learned is that there is a close and apparently reliable relationship between consumer spending, *C*, and disposable income, *DI*. Economists call this relationship the **consumption function.**

A second fact we have picked up from these figures is that the *slope* of the consumption function is fairly constant. We infer this from the fact that the straight line in Figure 24-4 comes so close to touching every point. If the slope of the consumption function had changed a lot, it would not be possible to do so well with a single straight line.[7] Because of its importance in such applications as the

[7]Figure 24-4 is limited to 17 years of data, but try fitting a single straight line to all the data in Figure 24-3. You will find that you can do remarkably well.

TABLE 24-1 CONSUMPTION AND INCOME IN MACROLAND

Year	(1) Consumption, C (billions of dollars)	(2) Disposable income, DI (billions of dollars)	(3) Marginal Propensity to Consume, MPC
1989	2,700	3,200	
			0.75
1990	3,000	3,600	
			0.75
1991	3,300	4,000	
			0.75
1992	3,600	4,400	
			0.75
1993	3,900	4,800	
			0.75
1994	4,200	5,200	

The **MARGINAL PROPENSITY TO CONSUME (MPC)** is the ratio of the change in consumption to the change in disposable income that produces the change in consumption. On a graph, it appears as the slope of the consumption function.

tax-cut example, economists have given a special name to this slope—the **marginal propensity to consume**, or **MPC** for short. The MPC tells us how much more consumers will spend if disposable income rises by $1.

$$\text{MPC} = \frac{\text{Change in C}}{\text{Change in } DI \text{ that produces the change in C}}$$

The MPC is best illustrated by an example, and for this purpose we turn away from U.S. data for a moment and look at the consumption and income data of a hypothetical country called Macroland (see Table 24-1). The data for Macroland resemble those for the United States, except that in Macroland, C and *DI* figures happen to be nice round numbers, which facilitates computation.

Columns 1 and 2 of Table 24-1 show annual consumer expenditure and disposable income from 1991 to 1996. These two columns constitute Macroland's consumption function and are plotted in Figure 24-5. Column 3 in the table shows the marginal propensity to consume (MPC), which is the slope of the line in Figure 24-5; it is derived from the first two columns. We can see that between 1993 and 1994, *DI* rose by $400 billion (from $4,000 billion to $4,400 billion) while C rose by $300 billion (from $3,300 billion to $3,600 billion). Thus, the MPC was:

$$\frac{\text{Change in C}}{\text{Change in } DI} = \frac{\$300}{\$400} = 0.75$$

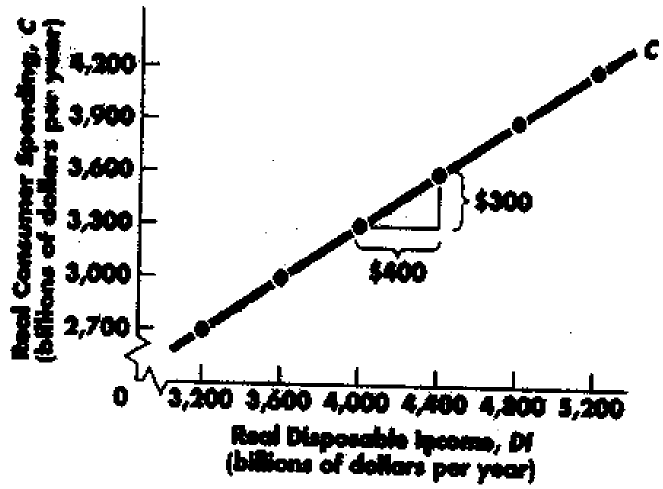
As you can easily verify, the MPC between any other pair of years in Macroland is also 0.75. This explains why the slope of the line in Figure 24-4 was so crucial in estimating the effect of a tax cut. This slope, which we found to be 0.90, is nothing but the MPC for the United States. It is the MPC that tells us how much *additional* spending will be induced by each dollar *change* in disposable income. For each $1 of tax cut, economists expect consumption to rise by $1 times the marginal propensity to consume.

To estimate the *initial* effect of a tax cut on consumer spending, economists must first estimate the MPC and then multiply the amount of the tax cut by the estimated MPC. But since they never know the true MPC with certainty, this prediction is always subject to some margin of error.[8]

[8]The word *initial* in the first sentence is an important one. Later chapters explain why the effects discussed in this chapter are only the beginning of the story.

FIGURE 24-5 THE CONSUMPTION FUNCTION OF MACROLAND

This diagram is similar to Figure 24-4, except that it applies to the hypothetical (and blissfully simple!) economy of Macroland. As can be seen, a straight-line consumption function passes through every point exactly. The slope of this line is 0.75, which is the marginal propensity to consume in Macroland.



In 1963, for example, economists multiplied the anticipated $9 billion tax cut by the estimated MPC of 0.90 and concluded that consumer spending would rise initially approximately $8 billion. Their estimate seems to have been quite accurate.

MOVEMENTS ALONG VERSUS SHIFTS OF THE CONSUMPTION FUNCTION

Unfortunately, this sort of calculation does not always yield such precise results. Among the most important reasons is that the consumption function does not always stand still. Sometimes it shifts.

You will recall from Chapter 4 the important distinction between a *movement along* a demand curve and a *shift* of the curve. A demand curve depicts the relationship between quantity demanded and *one* of its many determinants—price. Thus, a change in price causes a *movement along the demand curve*, but a change in any other factor that influences quantity demanded causes a *shift of the entire demand curve*.

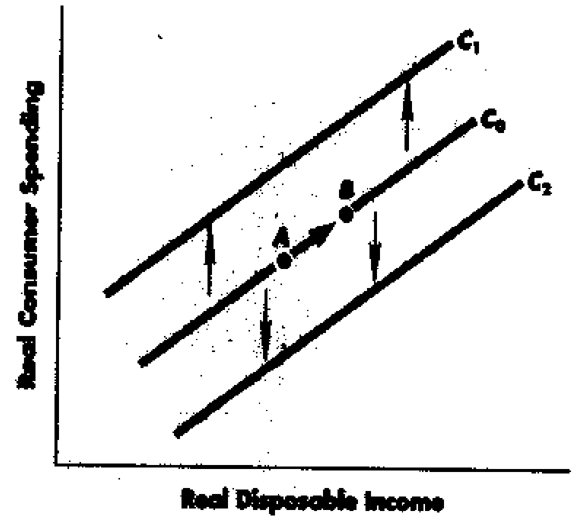
Because consumer spending is influenced by factors other than disposable income, a similar distinction is vital to understanding real-world consumption functions. Look back at the definition of the consumption function in the margin of page 571. A change in disposable income leads to a *movement along the consumption function* precisely because the consumption function depicts the relationship between C and *DI*. (See the orange arrow in Figure 24-6.) This is what we have been considering so far. But consumption also has other determinants, and a change in any of these will *shift the entire consumption function*—as indicated by the blue arrows in Figure 24-6. These shifts account for many of the errors in forecasting consumption. To summarize:

Any change in disposable income moves us *along* a given consumption function. But a change in any of the other determinants of consumption *shifts* the entire consumption schedule (see Figure 24-6).

Let us now list some of these "other determinants" that can shift the consumption function.

FIGURE 24-6 SHIFTS OF THE CONSUMPTION FUNCTION

An increase in disposable income causes a movement along a fixed consumption function, such as the movement from point *A* to point *B* on consumption function C_0 (see the orange arrow). But a change in any other determinant of consumer spending will cause the whole consumption function to shift upward (C_1) or downward (C_2).



OTHER DETERMINANTS OF CONSUMER SPENDING

WEALTH

One factor affecting consumption is consumers' *wealth,* which is a source of purchasing power in addition to income. Wealth and income are different things. For example, a wealthy person who does not work, but has a healthy bank account, may have little current *income.* Similarly, a high-income individual who spends all she earns will not accumulate wealth. To appreciate the importance of the distinction, consider two consumers, both earning $35,000 this year. One of them has $100,000 in the bank, while the other has no assets at all. Who do you think will spend more this year? Presumably the one with the big bank account.

The general point is that current income is not the only source of funds that households have; they can also finance spending by withdrawals from their bank accounts or by cashing in other forms of wealth. A stock market boom may therefore raise the consumption function (see the shift from C_0 to C_1 in Figure 24-6), while a collapse of stock prices may lower it (see the shift from C_0 to C_2).

THE PRICE LEVEL

A good deal of consumer wealth is held in forms whose values are fixed in money terms. Cash itself is the most obvious example of this, but government bonds, bank accounts, and corporate bonds are all assets with fixed face values in money terms. The purchasing power of any *money fixed asset* obviously declines whenever the price level rises, which means that the asset can buy less. For example, if the price level rises by 10 percent, a $1,000 government bond will buy about 10 percent less than it could when prices were lower. Consequently:

Higher prices decrease the demand for goods and services by eroding the purchasing power of consumer wealth.

This is no trivial matter. Consumers in the United States hold money fixed assets worth over $5 *trillion*, so that each 1 percent rise in the price level reduces the purchasing power of consumer wealth by over $50 billion, a tidy sum. The process, of course, operates equally well in reverse. Since a decline in the price level increases the purchasing power of money fixed assets:

Lower prices increase the demand for goods and services by enhancing the purchasing power of consumer wealth.

For these reasons, a change in the price level will shift the entire consumption function. Specifically:

A higher price level leads to lower real wealth and therefore to less spending *at any given level of real income*. Thus, a higher price level leads to a lower consumption function (such as C_2 in Figure 24-6). Conversely, a lower price level leads to a higher consumption function (such as C_1 in Figure 24-6).

Students often get confused on this point, so it is worth repeating that the depressing effect of the price level on consumer spending works through real *wealth*, not through real *income*. The consumption function is a relationship between *real consumer income* and *real consumer spending*. Thus, any decline in real income, regardless of its cause, moves the economy leftward *along a fixed consumption function*; it does not shift the consumption function.[9] By contrast, any decline in *real wealth* will *shift the whole consumption function downward*, meaning that there is less spending at any given level of real income.

THE INFLATION RATE

Prices may be high and rising slowly, or they may be low but rising rapidly. Therefore, the depressing effect of a high *price level* on real consumer spending must be distinguished from any effect on spending of the *rate of inflation* (that is, the rate at which prices are rising).

If inflation has any effect on consumer spending, it must be small. Economists are not even sure whether inflation stimulates or depresses spending. It used to be thought that high rates of inflation induce consumers to spend faster to "beat" the higher prices that loom on the horizon. But changes in consumer behavior when inflation rose in the 1970s and then fell in the 1980s belied this idea. Consumers have actually spent a *higher* fraction of their disposable incomes since the 1980s, when inflation was lower, than they did in the inflationary 1970s.

Because there is no strong evidence that the rate of inflation shifts the consumption function systematically in one direction or the other, we shall assume that the position of the consumption function is influenced by the *price level*, but not by the *inflation rate*.

THE REAL RATE OF INTEREST

A higher real rate of interest raises the rewards for saving. For this reason, many people believe it is "obvious" that higher real interest rates must encourage saving, and therefore discourage spending. Statistical studies of this relationship suggest otherwise, however. With very few exceptions, they show that interest

[9] This is true even if a rise in the price level lies behind the decline in real income. However, wages and prices normally move together, so there is no reason to expect real wages to fall when the price level rises.

POLICY DEBATE USING THE TAX CODE TO SPUR SAVING

Compared to the citizens of most industrial nations, Americans save rather little—only about 5 percent of disposable income in recent years. Many policymakers consider this a serious problem, so they have proposed numerous changes in the tax laws to increase incentives to save. During the 1995 session of Congress alone, for example, bills were discussed (but not passed) to do each of the following:

- Expand Individual Retirement Accounts ("IRAs"), which allow taxpayers to shelter some earnings from income tax by saving it
- Replace the income tax with a tax on consumer spending
- Change the current income tax into a "flat tax" that exempts interest and dividends from taxation

Each of these proposals, and others like them, would increase the after-tax return on saving. For example, under the current tax system, if you put away money at a 6 percent rate of interest and your income is taxed at a 30 percent rate, your after-tax rate of return on saving is just 4.2 percent (70 percent of 6 percent). But, if interest were tax-free, you would keep the full 6 percent. Members of Congress who advocate tax incentives for saving argue that they will induce Americans to save more.

The idea is eminently reasonable and has many supporters. Unfortunately, the evidence is squarely against it. Economists have conducted many studies of the effect of higher rates of return on saving. With very few exceptions, they detect little or no effect. Although the evidence fails to support the "common-sense" solution to the undersaving problem, the debate goes on. Many people, it seems, refuse to believe the evidence.

rates have virtually no effect on consumption decisions in the United States. Hence, in developing our model of the economy, we will assume that changes in interest rates do not shift the consumption function. (See the Policy Debate box, above.)

EXPECTATIONS OF FUTURE INCOMES

It is hardly earth shattering to suggest that consumers' expectations about future incomes may affect how much they spend today. This final determinant of consumer spending turns out to hold the key to resolving the puzzle posed at the beginning of the chapter: Why did tax policy succeed so well in 1964 and the early 1980s, but fail to alter consumer spending much in 1975?

WHY TAX POLICY FAILED IN 1975

To understand how expectations of future incomes affect current consumer expenditures, consider the abbreviated life histories of three consumers given in Table 24-2. (The reason for giving our three imaginary individuals such odd names will be apparent shortly.) The consumer named "No Change" earned $100 in each of the 4 years considered in the table. The consumer named "Temporary Rise" earned $100 in 3 of the 4 years, but had a good year in 1975. The consumer named "Permanent Rise" enjoyed a permanent increase in income in 1975 and was clearly the richest.

Now let us use our common sense to figure out how much each of these consumers might have spent in 1975. Temporary Rise and Permanent Rise had the same income that year. Do you think they spent the same amount? Not if they had some ability to foresee their future incomes, because Permanent Rise was richer in the long run.

Now compare No Change and Temporary Rise. Temporary Rise had 20 percent higher income in 1975 ($120 versus $100) but only 5 percent more over the entire 4-year period ($420 versus $400). Do you think his spending in 1975 was

TABLE 24-2 INCOMES OF THREE CONSUMERS

	Incomes in Each Year				
Consumer	**1974**	**1975**	**1976**	**1977**	**Total Income**
No change	100	100	100	100	**400**
Temporary rise	100	120	100	100	**420**
Permanent rise	100	120	120	120	**460**

closer to 20 percent above No Change's or closer to 5 percent above it? Most people guess the latter.

The point of this example is that it is reasonable for consumers to decide on their *current* consumption spending by looking at their *long-run* income prospects. This should come as no surprise to a college student. Are you spending only what you earn this year? Probably not. But that does not make you a foolish spendthrift. On the contrary, you know that your college education gives you a reasonable expectation of a much higher income in the future, and you are spending with that in mind.

Now let us see what all this has to do with the failure of the 1975 income tax cut. For this purpose, imagine that the three rows in Table 24-2 now represent the entire economy under three different government policies. Recall that 1975 was the year of the temporary tax cut. The first row (No Change) shows the unchanged path of disposable income if no tax cut was enacted. The second (Temporary Rise) shows an increase in disposable income attributable to a tax cut *for 1 year only*. The bottom row (Permanent Rise) shows a policy that increases *DI* in *every future year* by cutting taxes permanently in 1975. Which of the two lower rows do you imagine would have generated more consumer spending in 1975? The bottom row (Permanent Rise), of course. What we have concluded, then, is this:

Permanent cuts in income taxes cause greater increases in consumer spending than do temporary cuts of equal magnitude.

The application of this analysis to the case of the 1975 tax cut is immediate. The tax rebates were clearly one-time increases in income like that experienced by Temporary Rise in Table 24-2. No future income was affected, and so consumers did not increase their spending as much as government officials had hoped. The general lesson is:

A permanent increase in income taxes provides a greater deterrent to consumer spending than does a temporary increase of equal magnitude.

We have, then, what appears to be a general principle, backed up both by historical evidence and common sense. Permanent changes in income taxes have a more significant impact on consumer spending than do temporary changes. Though it may now seem obvious, this is not a lesson you would have learned from the introductory textbooks of 25 years ago. It is one that we learned the hard way, through bitter experience.

THE PREDICTABILITY OF CONSUMER BEHAVIOR

We have now learned enough to see why the economist's problem in predicting how consumers will react to an increase or decrease in taxes is not nearly as simple as suggested earlier in this chapter.

The principal problem seems to be anticipating how taxpayers will view any changes in the income tax law. If the government *says* that a tax cut is permanent, will consumers *believe* it and increase their spending accordingly? Perhaps not, if the government has a history of raising taxes after promising to keep them low. Similarly, when (as in 1975) the government explicitly announces that a tax cut is temporary, will consumers always believe this? Or might they greet such an announcement with a hefty dose of skepticism? This is quite possible if there is a history of "temporary" tax changes that stayed on the books indefinitely.

Thus, the effectiveness of any *future* tax policy move may well depend on the government's *past* track record. A government that repeatedly uses a succession of so-called "permanent" tax cuts and tax increases for short-run stabilization purposes may find consumers beginning to ignore the tax changes entirely. The story of the boy who cried wolf should probably be required reading for fiscal policy planners.

Nor is this the only problem. Consumer spending may be influenced by large and rapid accumulations of wealth (as happened immediately after World War II) or by sizable losses of wealth (such as the drastic decline in the stock market in 1987). Poor forecasts of future prices may lead consumption forecasts astray. Also, there are other hazards that we have not even mentioned here. For example, economists still do not understand why household saving fell (that is, why spending rose) so much in the mid-1980s. Economic predictions are inexact, and predictions of consumption illustrate this well.

There is much more that could be said about the determinants of consumption, but it is best to leave the rest to more advanced courses. For we are now ready to apply our knowledge of the consumption function to the construction of the first model of the whole economy. While it is true that income determines consumption, the consumption function in turn helps to determine the level of income. If that sounds like circular reasoning, read the next chapter!

SUMMARY

1. **Aggregate demand** is the total volume of goods and services purchased by consumers, businesses, government units, and foreigners. It can be expressed as the sum $C + I + G + (X - IM)$, where C is **consumer spending**, I is **investment spending**, G is **government purchases**, and $X - IM$ is **net exports**.

2. Aggregate demand is a schedule: the aggregate quantity demanded depends (among other things) on the price level. But, for any given price level, aggregate demand is a number.

3. Economists reserve the term *investment* to refer to purchases of newly produced factories, machinery, and houses.

4. Domestic product is the total volume of final goods and services produced in the country. It is most commonly measured by the gross domestic product.

5. **National income** is the sum of the *before-tax* wages, interest, rents, and profits earned by all individuals in the economy. By necessity, it must be approximately equal to domestic product.

6. **Disposable income** is the sum of the incomes of all individuals in the economy *after taxes and transfers*. It is the chief determinant of consumer expenditures.

7. All of these concepts, and others, can be depicted in a *circular flow diagram* that shows expenditures on all four sources flowing into business firms and national income flowing out.

8. The close relationship between consumer spending, C, and disposable income, *DI*, is called the **consumption function.** Its slope, which is used to predict the change in consumption that will be caused by a change in income taxes, is called the **marginal propensity to consume (MPC).**
9. Changes in disposable income move us *along a given consumption function*. Changes in any of the other variables that affect C *shift the entire consumption function*. Among the most important of these other variables are total consumer wealth, the price level, and expected future incomes.
10. Because consumers hold so many money fixed assets, they lose purchasing power when prices rise, which leads them to reduce their spending.
11. The government often has tried to manipulate aggregate demand by influencing private consumption decisions, usually through changes in the personal income tax. Although this policy seemed to work well in 1964 and 1981, it did not work well in 1975.
12. Future income prospects help explain these disparate events. The 1975 tax cut was temporary, and therefore left future incomes unaffected. By contrast, the 1964 and 1981–1984 tax cuts were permanent, and affected future as well as current incomes. It is no surprise, then, that the 1964 and 1981 actions had stronger effects on spending than did the 1975 action.

KEY TERMS

Aggregate demand
Consumer expenditure (C)
Investment spending (*I*)
Government purchases (G)
Net exports (*X* − *IM*)
C + *I* + G + (X − *IM*)
National income
Disposable income (*DI*)
Circular flow diagram
Transfer payments
Scatter diagram
Consumption function
Marginal propensity to consume (MPC)
Movements along versus shifts of the consumption function
Money fixed assets
Temporary versus permanent tax changes

QUESTIONS FOR REVIEW

1. What are the four components of aggregate demand? Which of these is the largest? Which is the smallest?
2. What is the difference between "investment" as the term is used by most people and "investment" as defined by an economist? Which of the following acts constitute investment according to the economist's definition?
   a. IBM opens a new factory to assemble personal computers.
   b. You buy 100 shares of IBM stock.
   c. A small computer company goes bankrupt, and IBM purchases its factory and equipment.
   d. Your family buys a newly constructed home from a developer.
   e. Your family buys an older home from another family. (*Hint:* Are any *new* products demanded by this action?)
3. What would the circular flow diagram (Figure 24-1, page 567) look like in an economy with no government? Draw one for yourself.
4. The marginal propensity to consume (MPC) for the nation as a whole is roughly 0.90. Explain in words what this means. What is your personal MPC?
5. Look at the scatter diagram in Figure 24-3 (page 570). What does it tell you about what was going on in this country in the years 1942 to 1945?
6. What is a consumption function, and why is it a useful device for government economists planning a tax cut?
7. On a piece of graph paper, construct the consumption function for Simpleland from the data given below and determine the MPC.

Year	Consumer Spending	Disposable Income
1991	1,200	1,500
1992	1,440	1,800
1993	1,680	2,100
1994	1,920	2,400
1995	2,160	2,700

8. In which direction will the consumption function for Simpleland shift if the price level rises? Show this on your graph.
9. Explain why permanent tax cuts are likely to lead to bigger increases in consumer spending than are temporary tax cuts.

10. (More difficult) Between 1990 and 1991, real disposable income (in 1992 dollars) rose only from $4,485 billion to $4,486 billion, owing to a recession. Use the data on real consumption expenditures given on the inside front cover of this book to compare the change in *C* to the change in *DI*. Explain why dividing the two does *not* give a good estimate of the marginal propensity to consume.

11. In 1995, several tax bills were debated in Congress that would have provided greater tax incentives for saving. (None were enacted.) If such saving incentives had been enacted, and had been successful, how would the consumption function have shifted?

APPENDIX NATIONAL INCOME ACCOUNTING

The type of macroeconomic analysis presented in this book dates from the publication of John Maynard Keynes's *The General Theory of Employment, Interest, and Money* in 1936. But at that time there was really no way to test Keynes's theories because the necessary data did not exist. It took some years for the theoretical notions used by Keynes to find concrete expression in real-world data.

The system of measurement devised for expressing data is called national income accounting.

The development of this system of accounts ranks as a great achievement in applied economics, perhaps as important in its own right as Keynes's theoretical work. Without it the practical value of Keynesian analysis would be severely limited. Economists spent long hours wrestling with the many difficult conceptual questions that arose in translating the theory into numbers, but they had one acknowledged leader: the late Professor Simon Kuznets, who was subsequently awarded the Nobel Prize in economics for his contributions to economic measurement techniques. Along the way some more-or-less arbitrary decisions and conventions had to be made. You may not agree with all of them, but the accounting framework that was devised is eminently serviceable, though, inevitably, it has some limitations that must be understood.

DEFINING GDP: EXCEPTIONS TO THE RULES

We first encountered the concept of gross domestic product (GDP) in Chapter 2.

Gross domestic product (GDP) is the sum of the money values of all final goods and services produced during a specified period of time, usually 1 year.

However, the definition of GDP has certain exceptions that we have not yet noted.

First, the treatment of government output involves a minor departure from the principle of using market prices. Outputs of private industries are sold on markets, so their prices are observed. But "outputs" of government offices are not sold; indeed, it is sometimes even difficult to define what those outputs are. Lacking prices for outputs, national income accountants fall back on the only prices they have: prices for the inputs from which the outputs are produced. Thus:

Government outputs are valued at the cost of the inputs needed to produce them.

This means, for example, that if a clerk at the Department of Motor Vehicles earns $10 an hour and spends one-half hour torturing you with explanations of why you cannot get a driver's license, that particular government "service" is considered as being worth $5, and it increases GDP by that amount.

Second, some goods that are produced during the year but not yet sold are nonetheless counted in that year's GDP. Specifically, goods that firms add to their *inventories* count in the GDP even though they do not pass through markets.

National income statisticians treat inventories as if they were "bought" by the firms that produced them, even though this "purchase" does not take place.

Finally, the treatment of investment goods runs slightly counter to the rule that only final goods are to be counted. In a broad sense, factories, generators, machine tools, and the like might be considered as intermediate goods. After all, their owners want them only for use in producing other goods,

not for any innate value that they possess. But this would present a real problem. Since factories and machines normally are never sold to consumers, when would we count them in GDP? National income statisticians avoid this problem by defining investment goods as final products demanded by the firms that buy them.

Now that we have a more complete definition of what the GDP is, let us turn to the problem of actually measuring it. National income accountants have devised three ways to perform this task, and we consider each in turn.

GDP AS THE SUM OF FINAL GOODS AND SERVICES

The first way to measure GDP seems to be the most natural, since it follows so directly from the circular flow diagram in this chapter. It also turns out to be the most useful definition for macroeconomic analysis. We simply add up the final demands of all consumers, business firms, government, and foreigners. Using the symbols Y, C, I, G, and $(X - IM)$ just as we did in the text, we have:

$$Y = C + I + G + (X - IM)$$

The I that appears in the actual U.S. national accounts is called **gross private domestic investment.**

The word *gross* will be explained presently. *Private* indicates that government investment is considered part of G, and *domestic* just means that machinery sold by American firms to foreign companies is included in exports rather than in I. Gross private domestic investment in the United States has three components: business investment in plant and equipment, residential construction (home building), and inventory investment. We repeat again that *only* these three things are *investment* in national income accounting terminology.

As defined in the national income accounts, *investment* includes only newly produced capital goods, such as machinery, factories, and new homes. It does not include exchanges of existing assets.

The symbol G, for government purchases, represents the *volume of current goods and services purchased by all levels of government*. Thus, anything that the government pays to its employees is counted in G, as are all of its purchases of goods.

Few citizens realize, however, that *most of what the federal government spends its money on is not for purchases of goods and services*. Instead, it is on *transfer payments*—literally, giving away money—either to individuals or to other levels of government.

The importance of the conceptual distinction lies in the fact that G represents the part of the national product that government uses up for its own purposes—to pay for armies, bureaucrats, paper, and ink—whereas transfer payments merely represent shuffling of purchasing power from one group of citizens to another group. Except for the administrators needed to run the programs, real economic resources are not used up in this process.

In adding up the nation's total output as the sum of $C + I + G + (X - IM)$, we are summing the shares of GDP that are used up by consumers, investors, government, and foreigners, respectively. Since transfer payments merely give someone the capability to spend on C, it is logical to exclude them from our definition of G, including in C only the portion of these transfer payments that is spent. If we included them in G, the same spending would get counted twice: once in G and then again in C.

The final component of GDP is net exports, which are simply exports of goods and services minus imports of goods and services. Table 24-3 shows GDP for 1995, in both nominal and real terms, computed as the sum of $C + I + G + (X - IM)$. You will note that the numbers for net exports in the table are actually negative. We will have much to say about America's trade deficit in Part VIII.

GDP AS THE SUM OF ALL FACTOR PAYMENTS

There is another way to count up the GDP—by *adding up all the incomes in the economy*. Let's see how this method handles some typical transactions. Suppose that General Electric builds a generator and sells it to General Motors for $1 million. The first method of calculating GDP simply counts the $1 million as part of I. The second method asks: What incomes resulted from the production of this generator? The answer might be something like this:

Wages of GE employees	$400,000
Interest to bondholders	50,000
Rentals of buildings	50,000
Profits of GE stockholders	100,000

TABLE 24-3

GROSS DOMESTIC PRODUCT IN 1995 AS THE SUM OF FINAL DEMANDS

Item	Amount (billions of current dollars)	Amount (billions of 1992 dollars)
Personal consumption expenditures (C)	4,924.3	4,578.5
Gross private domestic investment (I)	1,065.3	1,011.5
Government purchases of goods and services (G)	1,358.5	1,260.7
Net exports ($X - IM$)	−102.3	−114.2
Exports (X)	804.5	774.8
Imports (IM)	906.7	888.9
Gross domestic product (Y)	7,245.8	6,739.0

SOURCE: U.S. Department of Commerce. Totals do not add up precisely due to rounding.

The total is $600,000. The remaining $400,000 is accounted for by inputs that GE purchased from other companies: steel, circuitry, tubing, rubber, and so on.

But if we traced this $400,000 back further, we would find that it is accounted for by the wages, interest, and rentals paid by these other companies, *plus* their profits, *plus* their purchases from other firms. In fact, for *every* firm in the economy, there is an accounting identity that says:

Revenues from sales = Wages paid + Interest paid + Rentals paid + Profits earned + Purchases from other firms

Why must this always be true? Because profits are the balancing item; they are what is *left over* after the firm has made all its other payments. In fact, this accounting identity is really just a reorganization of the definition of profits: sales revenue less all costs of production.

Now apply this accounting identity to *all firms in the economy*. Total purchases from other firms are precisely what we call *intermediate goods*. What, then, do we get if we subtract these intermediate transactions from both sides of the equation?

Revenues from sales *minus* Purchases from other firms = Wages paid + Interest paid + Rentals paid + Profits earned

On the right-hand side, we have the sum of all factor incomes: payments to labor, land, and capital. On the left-hand side, we have total sales minus sales of intermediate goods. This means that we have only sales of *final* goods, which is precisely our definition of GDP. Thus, the accounting identity for the entire economy can be rewritten as:

GDP = Wages + Interest + Rents + Profits

and this gives national income accountants another way to measure the GDP.

Table 24-4 shows 1995's GDP measured by the sum of all incomes. Once again, we have omitted a few details in our discussion. The sum of wages, interest, rents, and profits actually adds up to only $5,799 billion (whereas GDP is $7,246 billion). We call this sum *national income* because it is the sum of all factor payments. But the actual selling prices of goods include another category of expense that we have ignored so far: sales taxes, excise taxes, and the like.

National income statisticians call these taxes *indirect business taxes*, and when we add them to national income, we obtain the **net national product (NNP).**

Notice here the use of the adjective *national* rather than *domestic*. When we add up all the wages, interest, rents, and profits received by Americans, we will inevitably include some payments derived from production in other countries. Similarly, some of the factor payments made by American businesses go to citizens of other countries.

If we subtract the former and add back the latter, we change net *national* product into **net domestic product (NDP).**

TABLE 24-4

GROSS DOMESTIC PRODUCT IN 1995 AS THE SUM OF INCOMES

Item	Amount (billions of dollars)	
Compensation of employees (wages)	4,209.1	
plus		
Net interest	401.0	
plus		
Rental income	122.2	
plus		
Profits	1,066.9	
Corporate profits		588.6
Proprietors' income		478.3
equals		
National income	5,799.2	
plus		
Indirect business taxes and miscellaneous items	612.4	
equals		
National net product	6,411.6	
minus		
income received from other countries	206.7	
plus		
income paid to other countries	215.0	
equals		
Net domestic product	6,419.9	
plus		
Depreciation	825.9	
equals		
Gross domestic product	7,245.8	

SOURCE: U.S. Department of Commerce. Totals do not add up precisely due to rounding.

Now we are almost at the GDP. The only difference between GDP and NDP is depreciation of the nation's capital stock.

Depreciation is the value of the portion of the nation's capital equipment that is used up within the year. It tells us how much output is needed just to keep the economy's capital stock intact.

The difference between gross and net simply refers to whether depreciation is included or excluded. We add depreciation to NDP to get GDP. Thus, GDP is a measure of all final output, taking no account of the capital used up in the process (and therefore in need of replacement). NDP deducts the required replacements to arrive at a *net* production figure.

From a conceptual point of view, most economists feel that NDP is a more meaningful indicator of the economy's output than GDP. After all, the depreciation component of GDP represents the output that is needed just to repair and replace worn-out factories and machines; it is not available for anybody to consume.[10] Therefore, NDP seems to be a better measure of well-being than GDP. But, alas, GDP is much easier to measure because depreciation is a particularly tricky item. What fraction of his tractor did Farmer Jones "use up" last year? How much did the Empire State Building depreciate during 1996? If you ask yourself these difficult questions, you will understand why most economists feel that GDP is measured more accurately

[10] If the capital stock is used for consumption, it will decline, and the nation will wind up poorer than before.

than is NDP. For this reason, most economic models are based on GDP.

In Table 24-4, you can hardly help noticing the preponderant share of employee compensation in total national income—almost 73 percent. Labor is by far the most important factor of production. The return on land is just 2 percent of national income, and interest accounts for about 7 percent. Profits account for the remaining 18 percent, though the size of corporate profits (less than 9 percent of GDP) is much less than the public thinks. If, by some magic stroke, we could eliminate all corporate profits without upsetting the performance of the economy, the average worker would get a raise of about 11 percent!

GDP AS THE SUM OF VALUES ADDED

It may strike you as strange that national income accountants include only *final* goods and services in GDP. Aren't *intermediate* goods part of the nation's product? They are, of course. The problem is that, if all intermediate goods were included in GDP, we would wind up double and triple counting things and therefore get an exaggerated impression of the actual level of economic activity.

To explain why, and to show how national income accountants cope with this difficulty, we must introduce a new concept, called *value added*.

The **value added** by a firm is its revenue from selling a product minus the amount paid for goods and services purchased from other firms.

The intuitive sense of the concept is clear: if a firm buys some inputs from other firms, does something to them, and sells the resulting product for a price higher than it paid for the inputs, we say that the firm has "added value" to the product. If we sum up the values added in this way by all the firms in the economy, we must get the total value of all final products. Thus:

GDP can be measured as the sum of the values added by all firms.

To verify this, look back at the second accounting identity on page 582. The left-hand side of this equation, sales revenue minus purchases from other firms, is precisely the firm's value added. Thus:

$$\text{Value added} = \text{Wages} + \text{Interest} + \text{Rents} + \text{Profits}$$

Since the second method we gave for measuring GDP is to add up wages, interest, rents, and profits, we see that the value-added approach must also yield the same answer.

The value-added concept is useful in avoiding double counting. Often, however, it is hard to distinguish intermediate goods from final goods. Paint bought by a painter, for example, is an intermediate good. But paint bought by a do-it-yourselfer is a final good. What happens, then, if the professional painter buys some paint to refurbish his own garage? The intermediate good becomes a final good. You can see that the line between intermediate goods and final goods is a fuzzy one in practice.

If we measure GDP by the sum of values added, however, it is not necessary to make such subtle distinctions. In this method, *every* purchase of a new good or service counts, but we do not count the entire selling price, only the portion that represents value added.

To illustrate this idea, consider the data in Table 24-5 and how they would affect GDP as the sum of final products. Our example begins when a farmer who grows soybeans sells them to a mill for \$3 a bushel. This transaction does *not* count in the GDP, because the miller does not purchase the soybeans

TABLE 24-5 AN ILLUSTRATION OF FINAL AND INTERMEDIATE GOODS

Item	Seller	Buyer	Price
Bushel of soybeans	Farmer	Miller	\$ 3
Bag of soy meal	Miller	Factory	4
Gallon of soy sauce	Factory	Restaurant	8
Gallon of soy sauce used as seasoning	Restaurant	Consumers	\$10
		Total:	\$25

Addendum: Contribution to GDP: \$10

TABLE 24-6 AN ILLUSTRATION OF VALUE ADDED

Item	Seller	Buyer	Price	Value Added
Bushel of soybeans	Farmer	Miller	$ 3	$ 3
Bag of soy meal	Miller	Factory	4	1
Gallon of soy sauce	Factory	Restaurant	8	4
Gallon of soy sauce used as seasoning	Restaurant	Consumers	10	2
			Total: $25	$10
	Addendum: Contribution to GDP			
	Final products	$10		
	Sum of values added	$10		

for his own use. The miller then grinds up the soybeans and sells the resulting bag of soy meal to a factory that produces soy sauce. The miller receives $4, but GDP still has not increased because the ground beans are also an intermediate product. Next, the factory turns the beans into soy sauce, which it sells to your favorite Chinese restaurant for $8. Still no effect on GDP.

But then the big moment arrives: The restaurant sells the sauce to you and other customers as a part of your meals, and you eat it. At this point, the $10 worth of soy sauce becomes a final product and is included in the GDP. Notice that if we had also counted the three intermediate transactions (farmer to miller, miller to factory, factory to restaurant), we would have come up with $25—2-1/2 times too much.

Why is it too much? The reason is straightforward. Neither the miller, the factory owner, nor the restaurateur values the product we have been considering *for its own sake*. Only the customers who eat the final product (the soy sauce) have had an increase in their material well-being, so only this last transaction counts in the GDP. However, as we shall now see, value-added calculations enable us to come up with the right answer ($10) by counting only *part* of each transaction. The basic idea is to count at each step only the contribution to the value of the ultimate final product that is made at that step, excluding the values of items produced at earlier steps.

Ignoring the minor items (such as fertilizer) that the farmer purchases from others, the entire $3 selling price of the bushel of soybeans is new output produced by the farmer; that is, the whole $3 is value added. The miller then grinds the beans and sells them for $4. He has added $4 minus $3, or $1 to the value of the beans. When the factory turns this soy meal into soy sauce and sells it for $8, it has added $8 minus $4, or $4 more in value. Finally, when the restaurant sells it to hungry customers for $10, a further $2 of value is added.

The last column of Table 24-6 shows this chain of creation of value added. We see that the total value added by all four firms is $10, exactly the same as the restaurant's selling price. This is as it must be, for only the restaurant sells the soybeans as a final product.

SUMMARY

1. Gross domestic product (GDP) is the sum of the money values of all final goods and services produced during a year and sold on organized markets. There are, however, certain exceptions to this definition.
2. One way to measure the GDP is to add up the final demands of consumers, investors, government, and foreigners: GDP = $C + I + G + (X - IM)$.
3. A second way to measure the GDP is to start with all the factor payments—wages, interest, rents, and profits—that constitute the national income and then add indirect business taxes and depreciation.
4. A third way to measure the GDP is to sum up the values added by every firm in the economy (and then once again add indirect business taxes and depreciation).
5. Except for possible bookkeeping and statistical errors, all three methods must give the same answer.

KEY TERMS

National income accounting
Gross domestic product (GDP)
Gross private domestic investment
Government purchases
Transfer payments
Net exports
National income
Net national product (NNP)
Net domestic product (NDP)
Depreciation
Value added
Personal income
Disposable income (*DI*)

QUESTIONS FOR REVIEW

1. Which of the following transactions are included in the gross domestic product, and by how much does each raise GDP?
   a. You buy a new Chevrolet, paying $12,000.
   b. You buy a new Toyota, imported from Japan, paying $17,000.
   c. You buy a used Chevrolet, paying $3,000.
   d. Ford builds a $100 million factory to make cars.
   e. Your grandmother receives a government Social Security check for $1,200.
   f. Chrysler manufactures 1,000 automobiles at a cost of $12,000 each. Unable to sell them, it holds them as inventories.
   g. Mr. Black and Mr. Blue, each out for a Sunday drive, have a collision in which their cars are destroyed. Black and Blue each hire a lawyer to sue the other, paying the lawyers $3,000 each for services rendered. The judge throws the case out of court.
   g. You sell a used refrigerator to your friend for $40.
2. Explain the difference between final goods and intermediate goods. Why is it sometimes difficult to apply this distinction in practice? In this regard, why is the concept of value added useful?
3. Explain the difference between government spending and government purchases of goods and services (G). Which is larger?
4. Explain why national income and gross domestic product would be essentially equal if there were no depreciation and no indirect business taxes.
5. The following outline provides a complete description of all economic activity in Trivialand for 1996. Draw up versions of Tables 24-3 and 24-4 for Trivialand showing GDP computed in two different ways.
   a. There are thousands of farmers but only two big business firms in Trivialand: Specific Motors (an auto company) and Super Duper (a chain of food markets). There is no government and no depreciation.
   b. Specific Motors produced 1,000 small cars, which it sold at $6,000 each, and 100 trucks, which it sold at $8,000 each. Consumers bought 800 of the cars, and the remaining 200 cars were exported to the United States. Super Duper bought all the trucks.
   c. Sales at Super Duper markets amounted to $14 million, all of it sold to consumers.
   d. All the farmers in Trivialand are self-employed and sell all their wares to Super Duper.
   e. The costs incurred by all the businesses were as follows:

	Specific Motors	Super Duper	Farmers
Wages	$3,800,000	$4,500,000	$2,000,000
Interest	100,000	200,000	700,000
Rent	200,000	1,000,000	2,000,000
Purchases of food	0	7,000,000	0

6. (More difficult) Now complicate Trivialand in the following ways and answer the same questions. In addition, calculate national income and disposable income.
   a. The government bought 50 cars, leaving only 150 cars for export. In addition, the government spent $800,000 on wages and it made $1,200,000 in transfer payments.
   b. Depreciation for the year amounted to $600,000 for Specific Motors and $200,000 for Super Duper. (The farmers had no depreciation.)
   c. The government levied sales taxes amounting to $500,000 on Specific Motors and $200,000 on Super Duper (none on farmers). In addition, the government levied a 10 percent income tax on all wages, interest, and rental income.
   d. In addition to the food and cars mentioned in Review Question 5, consumers in Trivialand imported 500 computers from the United States at $2,000 each.



CHAPTER 25

DEMAND-SIDE EQUILIBRIUM: UNEMPLOYMENT OR INFLATION?

Investment . . . is a flighty bird, which needs to be controlled.
J. R. Hicks

We learned in Chapter 22 that the interaction of aggregate demand and aggregate supply determines whether the economy will stagnate or prosper, whether our labor and capital resources will be fully employed or unemployed. And we learned in Chapter 24 that aggregate demand has four components: consumer expenditure (C), investment (I), government purchases (G), and net exports ($X - IM$). It is now time to start building a theory that puts all the pieces together.

Our approach is sequential. Since it is necessary to walk before you can run, we imagine in this chapter that the price level, the rate of interest, and the international value of the dollar are all constant. None of these assumptions are true, of course, and we will dispense with each later. But these three unrealistic assumptions enable us to construct a simple but useful model of how the state of aggregate demand influences the level of gross domestic product (GDP). In this simple model, only C is variable; the other three components of spending—I, G, and $X - IM$—are all assumed to be fixed.

Subsequent chapters will drop the three unrealistic assumptions in turn. In Chapter 27, we will bring in the supply side of the economy, which enables us to treat the price level as variable rather than constant. In Chapter 30, we will see how interest rates—and hence investment—are determined. Finally, Chapters 35 and 36 bring the exchange rate into the picture and study the determination of net exports.

But first things first. This chapter begins by examining the most volatile component of aggregate demand: investment. What factors determine investment spending, and why is it so variable and hard to predict?[1] Then we add net exports and

[1]We repeat the warning given in the previous chapter about the meaning of the word *investment*. It *includes* spending by businesses and individuals on *newly produced* factories, machinery, and houses. But it *excludes* sales of used industrial plants, equipment, and homes, and it *also excludes* purely financial transactions, such as the purchases of stocks and bonds.

government purchases to the model. Treating I, G, and $X - IM$ as constants, we next see how equilibrium is established on the demand side of the economy. Finally, we consider a question of great importance to policymakers: Can the economy be expected to achieve full employment of its resources if the government does not intervene?

PUZZLE: WHY DOES THE MARKET PERMIT UNEMPLOYMENT?

Economists are fond of pointing out—with some awe—the amazing achievements of free markets. Without central direction, they somehow get businesses to produce just the goods and services that consumers want—and to do so cheaply and efficiently. If consumers want less meat and more fish, markets respond; if people subsequently change their minds, markets respond again. Free markets seem to coordinate literally millions of decisions effortlessly and seamlessly.

Yet for hundreds of years and all over the globe, market economies have stumbled over one particular coordination problem: the periodic bouts of mass unemployment that we call *recessions* and *depressions*. Widespread unemployment represents a failure to coordinate economic activity in the following sense. If the unemployed were hired, they would be able to buy the goods and services that businesses cannot sell; revenues from those sales would, in turn, allow firms to pay the workers. So a seemingly straightforward "deal" offers jobs for the unemployed and sales for the firms. But somehow this deal is not consummated. Workers remain unemployed and firms get stuck with unsold output.

Thus, free markets, which somehow get rough diamonds dug out of the ground in South Africa and turned into beautiful rings that are bought for brides in Los Angeles, cannot seem to solve the coordination problem posed by unemployment. Why not? For centuries, economists puzzled over this question. By the end of the chapter, you will be well on the way toward providing an answer.

THE EXTREME VARIABILITY OF INVESTMENT

The first thing to be said about investment spending is that it is extraordinarily variable.

In Chapter 24, we learned that consumer spending follows movements in disposable income with great (though not perfect) reliability. Investment spending is quite different: It swings from high to low levels with astonishing speed. For example, when the U.S. economy declined a scant 1 percent (as measured by real GDP) between 1990 and 1991, real investment spending dropped a hefty 9.7 percent. Then, when the overall economy "boomed" at a 3.5 percent rate from 1993 to 1994, investment soared 14.3 percent. (See the box on the next page.) What accounts for these wide swings in investment spending?

BUSINESS CONFIDENCE AND EXPECTATIONS ABOUT THE FUTURE

While many factors influence business people's desires to invest, Keynes stressed the importance of the *state of business confidence*, which in turn depends on *expectations about the future*.

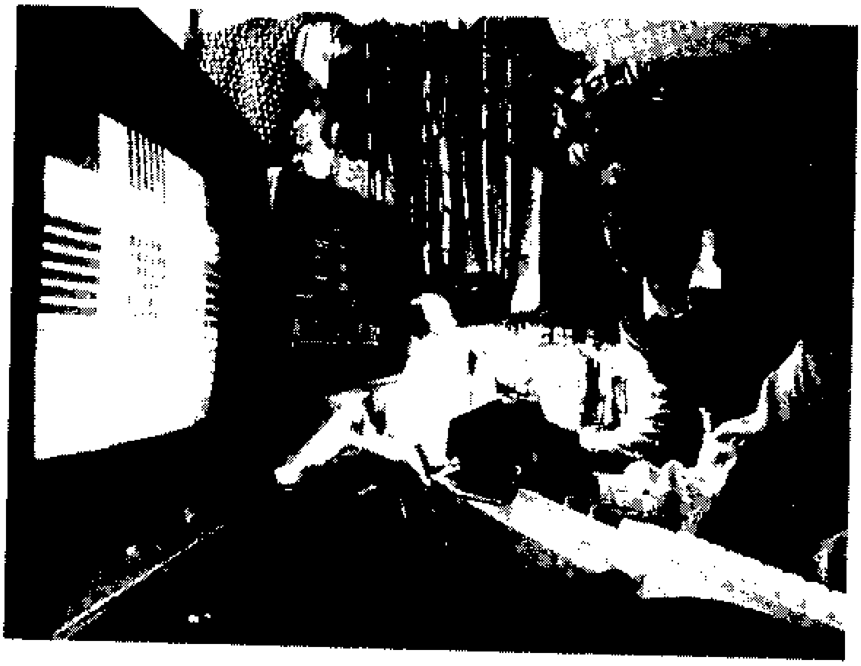
THE INVESTMENT BOOM OF THE NINETIES

[illegible] the U.S. economy emerged [illegible] the recession of 1990–1991, [illegible]ment spending has been rising at rates rarely seen since World War II. Between 1992 and 1994, outlays for such things as high-tech plants and computer-driven machinery jumped 20 percent. This surge in investment helped to boost productivity and fuel economic growth.

The last time investment spending rose at such a rapid pace was the 1960s, a time when the United States was the unquestioned powerhouse of the world economy. But back then, businesses were mainly looking for ways to increase production. Now the emphasis is on competing in the global market. Companies know they must efficiently create the highest-quality goods or services, create them on time, and exactly to customer specifications. If they do not, there are plenty of foreign and domestic firms ready to take their business away.

Technology is also fueling the boom. Although it may not be good news to some workers, many service companies have found that capital investments make more sense than maintaining large staffs. With modern equipment, factories and offices function better, faster, and more cheaply. Factories are being transformed through investments in new software and automation equipment. By embracing the latest technology, firms are becoming leaner and more efficient producers.

SOURCE: Joseph Spiers, "The Most Important Economic Event of the Decade," *Fortune*, April 3, 1995, p. 33.



While tricky to measure, it does seem obvious that businesses will build more factories and purchase more new machines when they are optimistic. Conversely, their investment plans will be very cautious if the economic outlook appears bleak. Keynes pointed out that psychological perceptions like these are subject to abrupt shifts, so that fluctuations in investment can be a major cause of instability in aggregate demand. Hence Hicks's analogy to a "flighty bird" in the chapter's opening quotation.

Unfortunately, neither economists nor, for that matter, psychologists have many good ideas about how to *measure*—much less *control*—business confidence. Therefore, economists usually focus on several more objective determinants of investment—determinants that are easier to quantify and, perhaps even more important, are more easily influenced by government policy.

THE LEVEL AND GROWTH OF DEMAND

Firms have strong incentives to invest when demand is pressing against capacity. Under these circumstances, business executives are likely to feel that new factories and machinery can be employed profitably. So, for example, firms like Intel and Motorola have recently been investing heavily in new plants to produce microchips. By contrast, if there is a great deal of unused machinery, empty

factories, and the like, managers may not find investment opportunities very attractive.

Because it takes a substantial amount of time to order machinery or to build a factory, *investment plans are made with an eye on the future.* Even when pressures on current capacity are not particularly severe, a firm that is expecting rapid growth in sales may start investing now in order to have adequate capacity for the future. Furthermore, briskly growing sales are likely to make business people more optimistic. Conversely, slow growth of output will discourage investment. In sum:

High levels of sales relative to current capacity and expectations of rapid economic growth create an atmosphere favorable to investment. Low levels of sales and slow anticipated growth are likely to discourage investment.

Government stabilization policy thus has an indirect handle on investment spending. By stimulating aggregate demand, it can induce business firms to invest more, though the precise amount may be hard to predict.

TECHNICAL CHANGE AND PRODUCT INNOVATION

Some investments are driven by technology. New investment opportunities suddenly appear when a new product like the mobile telephone is invented, or when a technological breakthrough makes an existing product much cheaper or better, as happened with microcomputers. In our capitalist market system, entrepreneurs seize these opportunities quickly, building new factories, stores, and offices. These new investments need not be "high tech." The VCR, for example, spawned an entire service industry of video rental shops that now dot the American landscape. Two decades ago, such stores did not even exist.

THE REAL RATE OF INTEREST

The real interest rate is the determinant of investment that will play the pivotal role in later chapters. When interest rates rise, investment falls, and it is not hard to see why. A good deal of business investment is financed by borrowing, and the interest rate indicates how much firms pay for that privilege. Some investment projects that yield a profit at an interest rate of 6 percent will be money-losers if the firm has to pay 10 percent.

The amount that businesses will want to invest depends on the real interest rate they must pay on their borrowings. The lower the real rate of interest, the more investment spending there will be.

In Chapter 30, we will study in some detail how the government can influence the rate of interest. Since interest rates affect investment, policymakers have another handle on aggregate demand—a handle they do not hesitate to use. The point is that—unlike business confidence, expectations, and technology—interest rates are visible and can be manipulated. Therefore, even if investment is much more sensitive to changes in confidence than to changes in interest rates, interest rates are nonetheless a more important instrument of government policy. But this is a topic for later in the book.

TAX PROVISIONS

The government has still another important way to influence investment spending—by altering various provisions of the tax law. For example, President Clin-

ton and the Republican Congress locked horns in 1995 over the advisability of reducing the tax rate on *capital gains*—the profit earned by selling an asset for more than you paid for it. Perhaps the principal argument on the Republican side was that lower capital gains taxes would lead to greater investment spending. But Democrats disputed that claim, and no changes were made.

In addition, there is a *tax on corporate profits,* which the government can reduce to spur investment—as it did in 1986. There are other, more complicated, tax provisions, as well. To summarize:

The tax law gives the government several ways to influence business spending on investment goods. But influence is far from control. Investment remains a "flighty bird."

THE DETERMINANTS OF NET EXPORTS

Another highly variable source of demand for U.S. products is foreign purchases of U.S. goods—our *exports.* However, as we learned in Chapter 24, to obtain the net contribution of foreigners to aggregate demand in the United States we must subtract *imports,* which is the portion of domestic demand that is satisfied by foreign producers.

NATIONAL INCOMES

While both exports and imports depend on many factors, the predominant one is *national income.* When consumption and investment spending by American consumers and firms rise, some of the increased spending is on foreign goods. Therefore:

Our imports rise when our GDP rises and fall when our GDP falls.

Similarly, our *exports* are the *imports* of other countries, so it is natural to assume that our exports depend on *their* GDPs, not on our own. Thus:

Our exports are relatively insensitive to our own GDP, but are quite sensitive to the GDPs of other countries.

Putting these two ideas together leads to a clear implication: When our economy grows faster than those of our trading partners, our net exports tend to shrink. Conversely, when foreign economies grow faster than ours, our net exports tend to rise.

RELATIVE PRICES AND EXCHANGE RATES

While GDP levels at home and abroad are important influences on a country's net exports, they are not the only relevant factors. International price differences matter, too. To make things concrete, let us focus on trade between the United States and Japan. Suppose that the prices of American goods rise while Japanese prices are constant. Then U.S. products become more expensive *relative to Japanese goods.* If American consumers react to the new relative prices by buying more Japanese goods, our *imports rise.* If Japanese consumers react to the same relative price changes by buying fewer American products, our *exports fall.* Both reactions reduce America's net exports.

Naturally, the effects of a decline in American prices are precisely the opposite: Exports are stimulated and imports are discouraged, so net exports rise. Thus:

A rise in the prices of a country's goods will lead to a reduction in that country's net exports. Analogously, a fall in the prices of a country's goods will raise that country's net exports.

Since trade patterns are governed by the prices of one country's goods *relative to* those of other countries, precisely the same logic applies to changes in Japanese prices. If Japanese prices fall while U.S. prices remain constant, Americans will import more and export less, so net exports $(X - IM)$ will decline. By similar reasoning, rising Japanese prices increase U.S. net exports. Thus:

Price increases abroad raise a country's net exports while price decreases abroad have the opposite effect.

This simple idea holds the key to understanding how rates of exchange among the world's currencies influence exports and imports—a topic we will consider in depth in Chapters 35 and 36. The reason is that exchange rates translate foreign prices into terms that customers are familiar with—their own currencies.

Consider, for example, Americans interested in buying British sweaters that cost £30. If the British pound is worth $1.50, the sweaters cost potential American buyers $45 each. But, if the pound is worth $2.00, those same sweaters cost Americans $60, and consumers are likely to buy fewer. These sorts of responses help explain why American auto makers won back market share from Japanese imports when the yen soared in the mid-1990s.

THE MEANING OF EQUILIBRIUM GDP

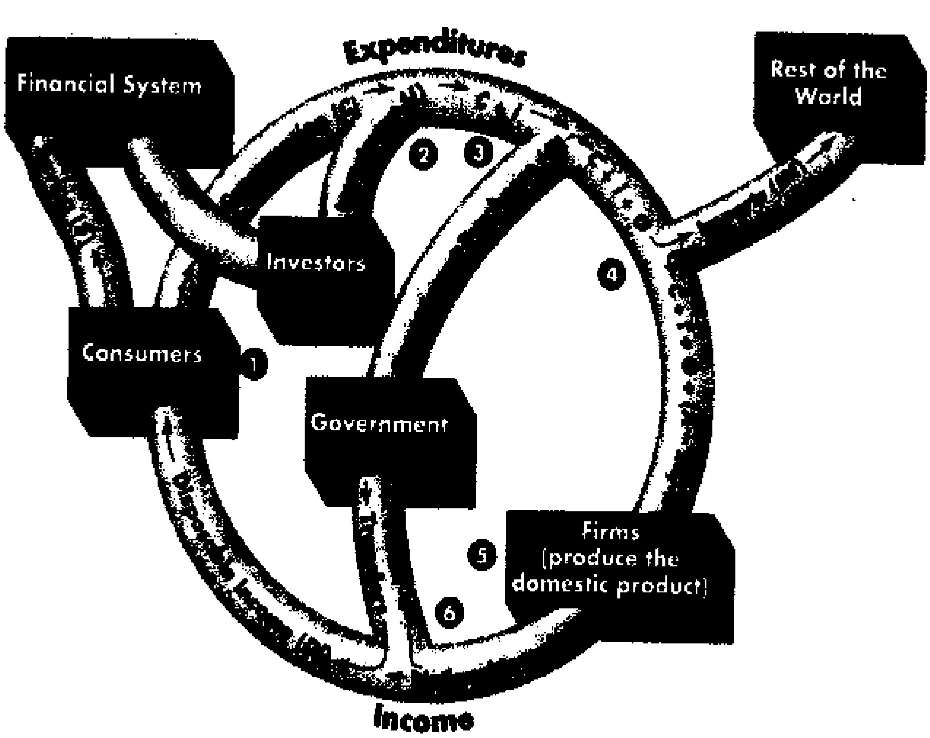
The preceding discussion accounts for three of the four main components of total spending. The fourth, government purchases of goods and services (G), is determined in the political arena by our elected representatives. Let us now put the four pieces together and see how they interact, using as our organizing framework the circular flow diagram introduced in the last chapter.

In doing so, we will at first ignore the possibility—raised in Chapter 22—that the government might vary its taxes (T) and spending (G) to steer the economy in some desired direction. Aside from pedagogical simplicity, there is an important reason for doing this. One of the crucial questions surrounding government stabilization policy is whether the economy would *automatically* gravitate toward full employment if the government would simply leave it alone. Keynes, contradicting the teachings of generations of economists before him, claimed that it would not. But Keynes's views remain controversial to this day. We can study the issue best by imagining an economy in which the government never tried to manipulate aggregate demand. This is just what we do in this chapter.

Look now at Figure 25-1, which repeats Figure 24-1 of the last chapter. We can use this circular flow diagram to begin the construction of a simple model of the determination of national income. But first we must understand what is meant by "equilibrium income."

As we explained in the last chapter, total *production* and total *income* must, of necessity, be equal. But the same need not be true of total *spending*. Imagine that, for some reason, the total expenditures, $C + I + G + (X - IM)$, being made after

FIGURE 25-1 THE CIRCULAR FLOW DIAGRAM



Here we repeat the circular flow of income and expenditures that we introduced in Chapter 24. Equilibrium occurs when $C + I + G + (X - IM)$ is equal to Y.

point 4 in the figure are greater than the value of the output being produced by the business firms at point 5.

Two things may happen in such a situation. Since consumers, businesses, government, and foreigners together are buying more than firms are producing, businesses are being forced to take goods out of their warehouses to meet customer demands. Thus, inventory stocks are falling. These inventory reductions are a signal to retailers of a need to increase their orders and to manufacturers of a need to step up their production. Consequently, production is likely to rise.

At some later date, if there is evidence that the high level of spending is not just a temporary aberration, either manufacturers or retailers (or both) may also respond to the buoyant sales performances by raising their prices. Economists therefore say that neither output nor the price level is in **equilibrium** when total spending exceeds the value of current production.

EQUILIBRIUM refers to a situation in which neither consumers nor firms have any incentive to change their behavior. They are content to continue with things as they are.

It is clear from the definition of *equilibrium* that the economy cannot be in equilibrium when total spending exceeds production, for the falling inventories demonstrate to firms that their production and pricing decisions were not quite appropriate.[2] Thus, since we normally use GDP to measure output:

The equilibrium level of GDP cannot be one at which total spending exceeds the value of output because firms will notice that inventory stocks are being depleted. They may first decide to increase production sufficiently to meet the higher demand. Later they may decide to raise prices, as well.

[2]All the models in this book assume, strictly for simplicity, that firms want constant inventories. Deliberate changes in inventories are treated in more advanced courses.

Now imagine the other case, in which the flow of spending reaching firms falls short of current production. Some output cannot be sold and winds up as additions to inventories. (This is just what happened to Nintendo and Sega when the demand for video games sagged in 1994.) The inventory pile-up acts as a signal to firms that at least one of their decisions was wrong. Once again, they will probably react first by cutting back on production, causing GDP to fall. If the imbalance persists, they may also lower prices to stimulate sales. But they certainly will not be happy with things as they are. Thus:

The equilibrium level of GDP cannot be one at which total spending is less than the value of output, because firms will not allow inventories to continue to pile up. They may decide to decrease production, or they may decide to cut prices in order to stimulate demand. Normally, firms are reluctant to cut prices until they are certain that the low level of demand is not a temporary phenomenon, so they rely more heavily on reductions in output.

EQUILIBRIUM ON THE DEMAND SIDE OF THE ECONOMY

We have now determined, through a process of elimination, the level of output that is consistent with people's desires to spend. We have reasoned that GDP will rise whenever it is below total spending, $C + I + G + (X - IM)$, and that GDP will fall whenever it is above $C + I + G + (X - IM)$. Equilibrium can occur, then, only when there is just enough spending to absorb the current level of production. Under such circumstances, producers conclude that their price and output decisions are correct. They therefore have no incentive to change those decisions. We conclude that:

The *equilibrium level of GDP on the demand side* is the one at which total spending equals production. In such a situation, firms find their inventories remaining at desired levels, so there is no incentive to change output or prices.

Thus, the circular flow diagram has helped us to understand the concept of equilibrium GDP on the demand side. It has also shown us how the economy is driven toward this equilibrium. It leaves unanswered, however, three important questions:

1. How large is the equilibrium level of GDP?
2. Will the economy suffer from unemployment, inflation, or both?
3. Is the equilibrium level of GDP on the demand side also consistent with firms' desires to produce? That is, is it also an equilibrium on the *supply* side?

The first two questions will occupy our attention in this chapter; the third question is reserved until Chapter 27.

CONSTRUCTING THE EXPENDITURE SCHEDULE

Our first objective is to determine precisely the equilibrium level of GDP and to see what factors it depends upon. To make the analysis more concrete, we turn to a numerical example. Specifically, we examine the relationship between total spending and GDP in Macroland, the hypothetical economy that was introduced in the last chapter.

TABLE 25-1 TOTAL EXPENDITURE IN MACROLAND (BILLIONS OF DOLLARS)

(1) GDP (Y)	(2) Consumption (C)	(3) Investment (I)	(4) Government Purchases (G)	(5) Net Exports (X − IM)	(6) Total Expenditure
4,800	3,000	900	1,300	−100	5,100
5,200	3,300	900	1,300	−100	5,400
5,600	3,600	900	1,300	−100	5,700
6,000	3,900	900	1,300	−100	6,000
6,400	4,200	900	1,300	−100	6,300
6,800	4,500	900	1,300	−100	6,600
7,200	4,800	900	1,300	−100	6,900

Columns 1 and 2 of Table 25-1 repeat the consumption function of Macroland that we first encountered in Table 24-1. They show how consumer spending, C, depends on GDP, which we symbolize by the letter Y. Columns 3 to 5 provide the other three components of total spending, I, G, and $X - IM$, through the simplifying assumptions that each is just a fixed number regardless of the level of GDP. Specifically, we assume that investment spending is \$900 billion, government purchases are \$1,300 billion, and net exports are −\$100 billion—meaning that in Macroland, as in the United States at present, imports exceed exports.

By adding together Columns 2 through 5, we calculate $C + I + G + (X - IM)$, or total expenditure, which is displayed in Column 6. Columns 1 and 6 are shaded to highlight how total expenditure depends on income in Macroland. We call this relationship the **expenditure schedule.**

An **EXPENDITURE SCHEDULE** shows the relationship between national income (GDP) and total spending.

Figure 25-2 shows the construction of the expenditure schedule graphically. The black line labeled C is the consumption function of Macroland. It plots on a graph the numbers given in Columns 1 and 2 of Table 25-1.

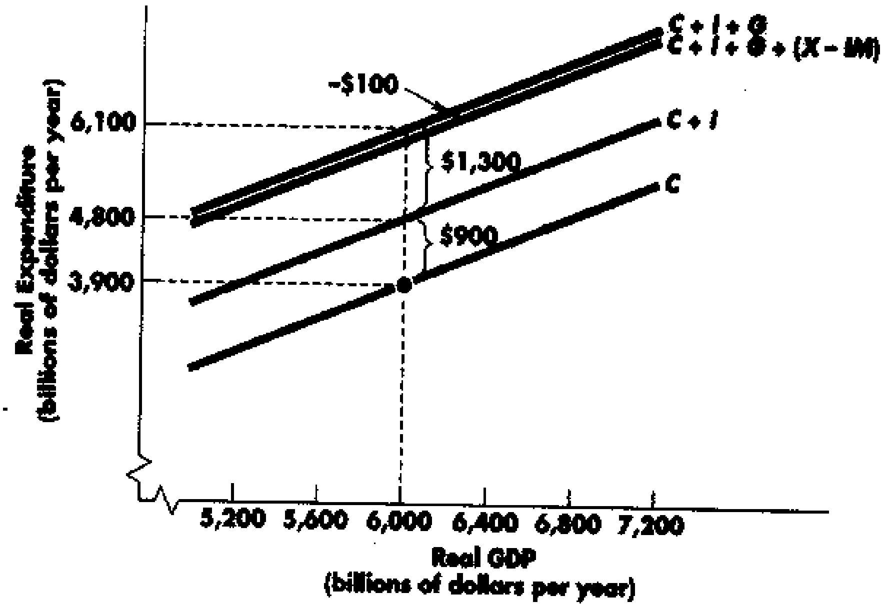
The orange line, labeled $C + I$, displays our assumption that investment is fixed at \$900 billion, regardless of the level of GDP. It lies a fixed distance (corresponding to \$900 billion) above the C line. If investment were not always \$900 billion, the two lines would either move closer together (at income levels at which investment was below \$900 billion) or grow farther apart (at income levels at which investment was above \$900 billion). For example, our list of determinants of investment spending suggested that I might be larger at higher levels of GDP. Because of this added investment—which is called **induced investment**—the resulting $C + I$ schedule would have a steeper slope than the C schedule.

INDUCED INVESTMENT is the part of investment spending that rises when GDP rises and falls when GDP falls.

The blue line, labeled $C + I + G$, adds in government purchases. Since they are assumed to be \$1,300 billion regardless of the size of GDP, the blue line is parallel to the orange line and \$1,300 billion higher.

Finally, the burgundy line labeled $C + I + G + (X - IM)$ adds in net exports. It is parallel to the blue line and \$100 billion lower, reflecting our assumption that net exports in Macroland are always −\$100 billion. Once again, if imports depended on GDP, as our previous discussion suggested, the $C + I + G$ and $C + I + G + (X - IM)$ lines will not be parallel.

FIGURE 25-2 CONSTRUCTION OF THE EXPENDITURE SCHEDULE



This figure shows in a diagram what Table 25-1 showed numerically—the construction of a total expenditure schedule from its components. Line *C* is the consumption function that we first encountered in Figure 24-5. Line *C* + *I* adds investment (assumed always to be $900 billion in this example), and line *C* + *I* + *G* adds government purchases (which are $1,300 billion). Line *C* + *I* + *G* + (*X* – *IM*) is the expenditure schedule and is obtained by adding net exports to *C* + *I* + *G*. For example, when GDP is $6,000, *C* is $3,900, *I* is $900, *G* is $1,300, and (*X* – *IM*) is –$100, for a total of $6,000.

THE MECHANICS OF INCOME DETERMINATION

We are now ready to determine demand-side equilibrium in Macroland. Look first at Table 25-2, which presents the logic of our circular flow argument in tabular form. The first two columns of this table reproduce the expenditure schedule that was constructed in Table 25-1. The other columns explain the process by which equilibrium is approached. Let us see why a GDP of $6,000 billion must be the equilibrium level.

Consider first any output level below $6,000 billion. For example, at output level Y = $5,200 billion, total expenditure is $5,400 billion (Column 2), which is $200 billion more than production. With spending greater than output (Column 3), inventories will be falling (Column 4). As the table suggests, this will be a signal to producers to raise their output (Column 5). Clearly, then, no output level below Y = $6,000 billion can be an equilibrium. Output is too low.

A similar line of reasoning can eliminate any output level above $6,000 billion. Consider, for example, Y = $6,800 billion. The table shows that total spending would be $6,600 billion if national income were $6,800 billion, so $200 billion of the GDP would go unsold. This would raise producers' inventory stocks and signal them that their rate of production was too high.

Just as we concluded from our circular flow diagram, equilibrium will be achieved only when total spending, $C + I + G + (X - IM)$, is equal to GDP (Y). In symbols, our condition for equilibrium GDP is:

$$Y = C + I + G + (X - IM)$$

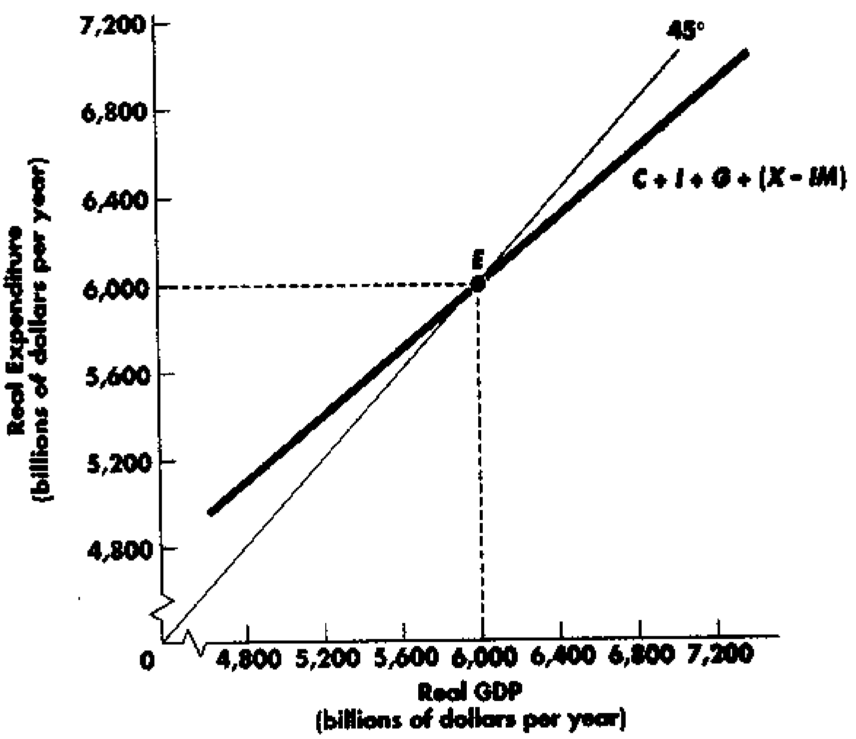
The table shows that this occurs only at a GDP of $6,000 billion. This, then, must be the equilibrium level of GDP.

TABLE 25-2 THE DETERMINATION OF EQUILIBRIUM OUTPUT

(1) Output (Y) (billions of dollars)	(2) Total Spending [C + I + G + (X − IM)] (billions of dollars)	(3) Balance of Spending and Output	(4) Inventories Are:	(5) Producers Will Respond by:
4,800	5,100	Spending exceeds output	Falling	Producing more
5,200	5,400	Spending exceeds output	Falling	Producing more
5,600	5,700	Spending exceeds output	Falling	Producing more
6,000	6,000	Spending = output	Constant	Not changing production
6,400	6,300	Output exceeds spending	Rising	Producing less
6,800	6,600	Output exceeds spending	Rising	Producing less
7,200	6,900	Output exceeds spending	Rising	Producing less

Figure 25-3 shows this same conclusion graphically, by adding a 45° line to Figure 25-2. Why a 45° line? Recall from the appendix to Chapter 1 that a 45° line marks all points on a graph at which the value of the variable measured on the horizontal axis is equal to the value of the variable measured on the vertical axis. In this convenient graph of the expenditure schedule, gross domestic product (Y) is measured on the horizontal axis and total expenditure, $C + I + G + (X - IM)$, is measured on the vertical axis. So the 45° line shows all the points at which output and spending are equal; that is, where $Y = C + I + G + (X - IM)$. *The 45° line therefore displays all the points at which the economy can possibly be at equilibrium*, for if total spending is not equal to production, firms will not be content with current output levels.

FIGURE 25-3 INCOME-EXPENDITURE DIAGRAM



This figure adds a 45° line—which marks off points where expenditure and output are equal—to Figure 25-2. Since the condition for equilibrium GDP is that expenditure and output must be equal, this line can be used to determine the equilibrium level of GDP. In this example, equilibrium is at point E, where GDP is \$6,000 billion—precisely as we found in Table 25-2.

Now we must compare these *potential* equilibrium points with the *actual* combinations of spending and output that the economy can attain, given the behavior of consumers and investors. That behavior, as we have seen, is described by the $C + I + G + (X - IM)$ line in Figure 25-3, which shows how total expenditure varies as income changes. Thus, *the economy will always be on the expenditure line* because only points on the $C + I + G + (X - IM)$ line are consistent with the spending plans of consumers and investors. Similarly, *if* the economy is in equilibrium, it *must* be on the 45° line. As Figure 25-3 shows, these two requirements together imply that the only viable equilibrium is at point E, where the $C + I + G + (X - IM)$ line intersects the 45° line. Only this point is consistent both with equilibrium and with the actual desires to consume and invest.

Notice that to the left of the equilibrium point, E, the expenditure line lies above the 45° line. This means that total spending exceeds total output, as we have already noted in words and with numbers. Hence, inventories will be falling and firms will conclude that they should increase production. Thus, production will rise toward the equilibrium point, E. The opposite is true to the right of point E. Here spending falls short of output, inventories are rising, and firms will cut back production—thereby moving closer to E.

An **INCOME-EXPENDITURE DIAGRAM,** also called a **45° LINE DIAGRAM,** plots total real expenditure (on the vertical axis) against real income (on the horizontal axis). The 45° line marks off points where income and expenditure are equal.

In other words, whenever production is above the equilibrium level, market forces will drive output down. And whenever production is below equilibrium, market forces will drive output up. Thus, in either case, deviations from equilibrium will be eliminated.

Diagrams like this one will recur so frequently in this and the next several chapters that it will be convenient to have a name for them. Let us therefore call them **income-expenditure diagrams** since they show how expenditures vary with income. Sometimes we shall also refer to them simply as **45° line diagrams.**

THE AGGREGATE DEMAND CURVE

Chapter 22 sketched a framework for macroeconomic analysis by introducing aggregate demand and aggregate supply curves, which relate aggregate quantities demanded and supplied to the price level. The expenditure schedule graphed in Figure 25-3 is certainly not the aggregate demand curve, for we have yet to bring the price level into our discussion. It is now time to remedy this omission and derive the aggregate demand curve.

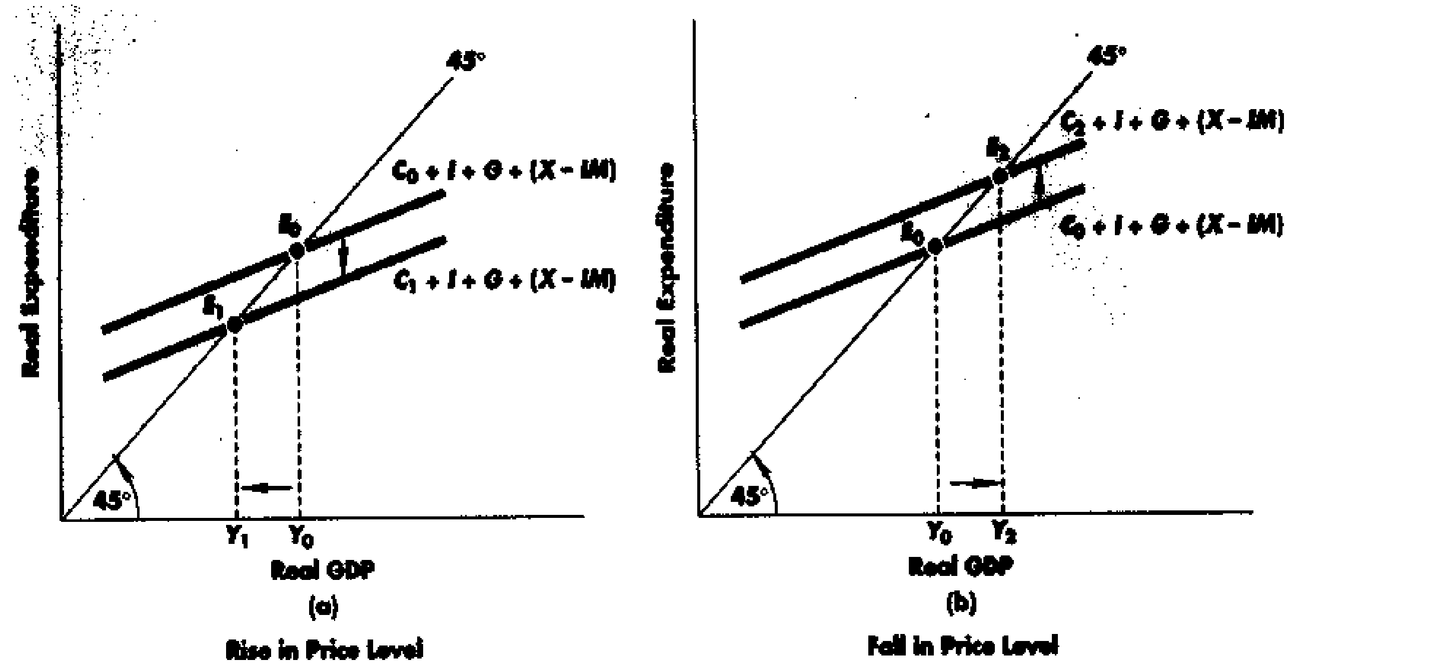
Fortunately, no further mechanical apparatus is required. We can bring the price level into our income-expenditure analysis by recalling something we learned in the last chapter: At any given level of real income, higher prices lead to lower real consumer spending. The reason, you will recall, is that consumers own many assets whose values are fixed in money terms, and which therefore lose purchasing power when prices rise.[3] With real wealth lower, consumers spend less. Therefore, total spending in the economy falls *even with no change in real income.*

In terms of our 45° line diagram, a rise in the price level will pull down the consumption function depicted in Figure 25-2 and, hence, will pull down the total expenditure schedule, as well. Conversely, a fall in the price level will raise both the C and $C + I + G + (X - IM)$ schedules in the diagram. The two parts of Figure 25-4 illustrate both of these shifts.

---

[3]The money in your bank account is a prime example. If prices rise, it will buy less.

FIGURE 25-4 THE EFFECT OF THE PRICE LEVEL ON EQUILIBRIUM AGGREGATE QUANTITY DEMANDED



Because a change in the price level causes the expenditure schedule to shift, it changes the equilibrium quantity of real GDP demanded. Panel (a) shows what happens when the price level rises, causing the expenditure schedule to shift downward from $C_0 + I + G + (X - IM)$ to $C_1 + I + G + (X - IM)$. Equilibrium quantity demanded falls from Y_0 to Y_1. Panel (b) shows what happens when the price level falls, causing the expenditure schedule to shift upward from $C_0 + I + G + (X - IM)$ to $C_2 + I + G + (X - IM)$. Equilibrium quantity demanded rises from Y_0 to Y_2.

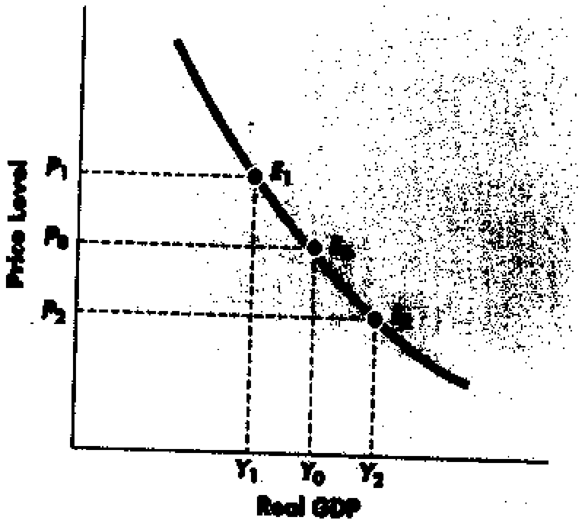
What, then, do changes in the price level do to the equilibrium level of real GDP on the demand side? Common sense says that, with lower spending, equilibrium GDP should fall, and Figure 25-4 shows that this conclusion is correct. Panel (a) shows that a rise in the price level, by shifting the expenditure schedule downward from $C_0 + I + G + (X - IM)$ to $C_1 + I + G + (X - IM)$ leads to a reduction in the equilibrium quantity of real GDP demanded from Y_0 to Y_1. Panel (b) shows that a fall in the price level, by shifting the expenditure schedule upward from $C_0 + I + G + (X - IM)$ to $C_2 + I + G + (X - IM)$, leads to a rise in the equilibrium quantity of real GDP demanded from Y_0 to Y_2. In summary:

A rise in the price level leads to a lower equilibrium level of real aggregate quantity demanded. This relationship between the price level and the equilibrium quantity of real GDP demanded is depicted in Figure 25-5 on the next page and is precisely what we called the *aggregate demand curve* in earlier chapters. It comes directly from the 45° line diagrams in Figure 25-4. Thus, points E_0, E_1, and E_2 in Figure 25-5 correspond precisely to the points bearing the same labels in Figure 25-4.

The effect of higher prices on consumer wealth is just one of several reasons why the aggregate demand curve relating the price level to real GDP demanded slopes downward. A second reason comes from international trade. In our discussion of the determinants of net exports (see pages 591–592), we pointed out that higher U.S. prices will depress exports (X) and stimulate imports (IM), provided that foreign prices are held constant. That means that, other things equal, a higher U.S. price level will reduce the $(X - IM)$ component of total

FIGURE 25-5 THE AGGREGATE DEMAND CURVE

The graphic analysis in Figure 25-4 showed that higher prices lead to lower aggregate quantity demanded. This relationship is called the *aggregate demand curve* and is shown in this figure.



expenditure, thereby shifting the $C + I + G + (X - IM)$ line downward and lowering real GDP, as depicted in Figure 25-4(a).

Later in the book, after we have studied interest rates and exchange rates, we will encounter still more reasons for a downward-sloping aggregate demand curve. All of them imply that:

An income-expenditure diagram like Figure 25-3 can be drawn up only for a *specific* price level. At different price levels, the $C + I + G + (X - IM)$ schedule will be different and, hence, the equilibrium quantity of GDP demanded will be different.

As we shall now see, this finding is critical to understanding the genesis of unemployment and inflation.

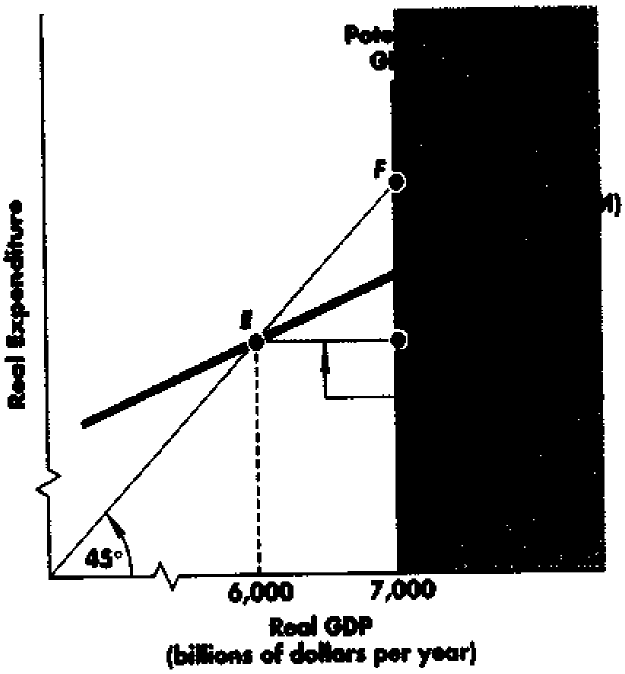
DEMAND-SIDE EQUILIBRIUM AND FULL EMPLOYMENT

We now turn to the second major question of this chapter: Will the economy achieve an equilibrium at full employment without inflation, or will there be unemployment, inflation, or both? This is one of the crucial questions surrounding government stabilization policy, for if the economy always gravitates toward full employment *automatically*, then the government should simply leave it alone.

In the income-expenditure diagrams used so far, the equilibrium level of GDP demanded has been shown as the intersection of the expenditure schedule and the 45° line, regardless of whatever level of GDP might correspond to full employment. However, as we will see now, when equilibrium GDP falls above full employment, the economy probably will be plagued by inflation, and when equilibrium falls below full employment, there will be unemployment and recession.

This remarkable fact was one of the principal messages of Keynes's *General Theory of Employment, Interest, and Money*. Writing during the Great Depression, it was natural for him to focus on the case in which equilibrium falls short of full employment so that there are unemployed resources. Figure 25-6 illustrates

FIGURE 25-6 A RECESSIONARY GAP



Sometimes equilibrium GDP may fall below potential GDP, so that some workers are unemployed. This diagram illustrates such a case. The horizontal distance *EB* between equilibrium GDP and potential GDP is called the *recessionary gap*.

this possibility. A vertical line has been drawn at the full-employment level of GDP (called *potential GDP*), which is assumed to be $7,000 billion in the example. We see that the $C + I + G + (X - IM)$ curve cuts the 45° line at point E, which corresponds to a GDP (Y = $6,000 billion) below potential GDP. In this case, the expenditure curve is too low to lead to full employment.

Such a situation might arise because consumers or investors are unwilling to spend at normal rates, because government spending is low, because foreign demand is weak, or because the price level is "too high." Any of these would depress the $C + I + G + (X - IM)$ curve. Unemployment must then occur because not enough output is demanded to keep the entire labor force at work.

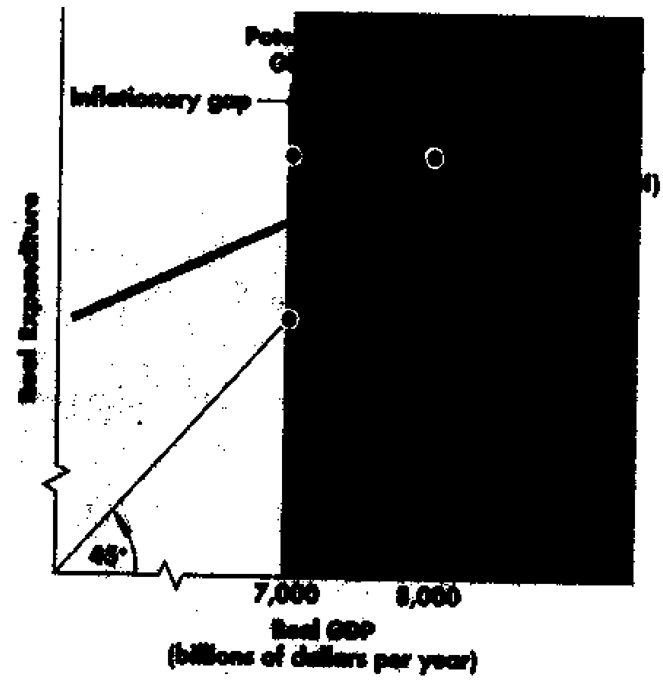
The **RECESSIONARY GAP** is the amount by which the equilibrium level of real GDP falls short of potential GDP.

The distance between the *equilibrium* level of output demanded and the *full-employment* level of output (that is, potential GDP) is called the **recessionary gap**—and is shown by the horizontal distance from E to B. While Figure 25-6 is entirely hypothetical, real-world gaps of precisely this sort were shown shaded blue in Figure 23-2 (page 541). They are a pervasive feature of recent U.S. economic history.

It is clear from Figure 25-6 that full employment can be reached only by raising the total spending schedule to eliminate the recessionary gap. Specifically, the $C + I + G + (X - IM)$ schedule must move upward until it cuts the 45° line at point F. Can this happen without government intervention? We know that a sufficiently large drop in the price level can do the job. But is that a realistic prospect? We shall return to this question after we bring the supply side into the picture. But first let us consider the other case, in which equilibrium GDP exceeds full employment.

Figure 25-7 illustrates this possibility. Now the expenditure schedule intersects the 45° line at point E, where GDP is $8,000 billion. But this exceeds the full-employment level, Y = $7,000 billion. A case like this can arise when consumer or investment spending is unusually buoyant, when foreign demand is

FIGURE 25-7 AN INFLATIONARY GAP



Sometimes equilibrium GDP may lie above potential GDP, meaning that there are more jobs than required for full employment. This diagram illustrates such a case. The horizontal distance *BE* between potential GDP and equilibrium GDP is called the *inflationary gap*. It is gradually eliminated by rising prices, which pull the $C + I + G + (X - IM)$ schedule down until it passes through point *F*.

particularly strong, when the government spends too much, or when a "low" price level pushes the $C + I + G + (X - IM)$ curve upward.

To reach an equilibrium at full employment, the price level would have to rise enough to drive the expenditure schedule *down* until it passed through point *F*. The horizontal distance *BE*—which indicates the amount by which the quantity of GDP demanded exceeds potential GDP—is called the **inflationary gap.** If there is an inflationary gap, a higher price level or some other means of reducing total expenditure is necessary to reach an equilibrium at full employment. Real-world inflationary gaps were shown shaded orange in Figure 23-2.

The **INFLATIONARY GAP** is the amount by which equilibrium real GDP exceeds the full-employment level of GDP.

In sum, only if the price level and spending plans are "just right" will the expenditure curve intersect the 45° line precisely at full employment, so that neither a recessionary gap nor an inflationary gap occurs. Are there reasons to expect this outcome? Does the economy have a self-correcting mechanism that automatically eliminates recessionary or inflationary gaps and propels it toward full employment? And how is it that inflation and unemployment sometimes occur together?

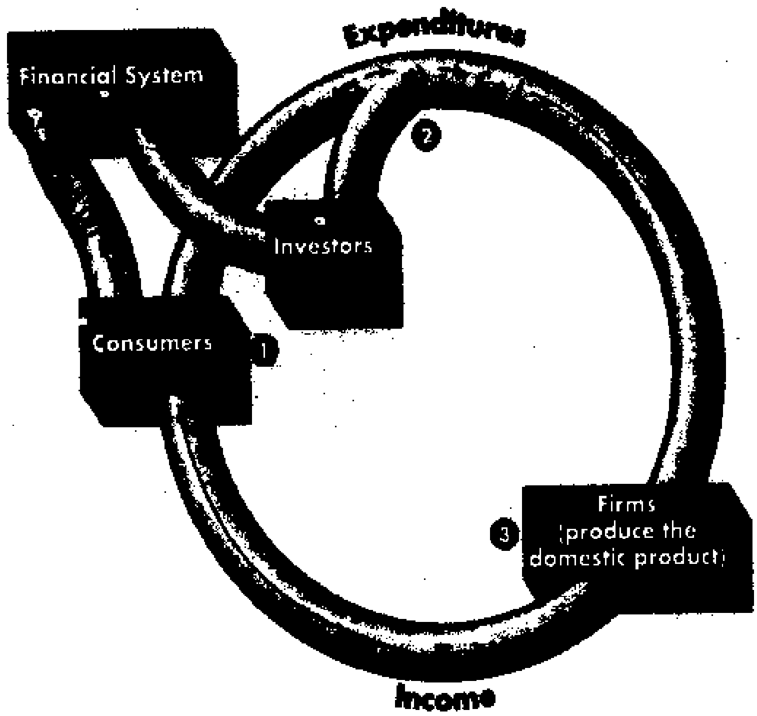
These are questions that we are not ready to address because we have not yet brought *aggregate supply* into the picture. And as we learned in Chapter 22, the price level is determined by the interaction of *both* aggregate demand *and* aggregate supply. However, it is not too early to get an idea about why things can go wrong, why the economy can find itself far away from full employment.

THE COORDINATION OF SAVING AND INVESTMENT

To understand what goes wrong in a recession, it is useful to pose the following question: Must the full-employment level of GDP be an equilibrium? Decades

FIGURE 25-8 A SIMPLIFIED CIRCULAR FLOW

Here we show a simplified version of the circular flow of income and expenditures shown in Figure 25-1. The simplification amounts to shutting off the pipes leading into and out of the government and those into and out of the rest of the world. Thus, this circular flow represents an economy with no government and no foreign trade.



ago, economists thought the answer was *yes*. Since Keynes, most economists believe the answer is *not necessarily*.

To help us understand why, Figure 25-8 offers a simplified circular flow diagram that ignores exports, imports, and the government. In this version, there is just one place for income to "leak out" of the circular flow: at point 1, where consumers save some of their income. Similarly, there is just one place for this lost spending to be replaced: at point 2, where investment enters the circular flow.

What happens if firms produce exactly the full-employment level of GDP at point 3 in the diagram? Will this income level be maintained as we move around the circle, or will it shrink or grow? The answer is that full-employment income will be maintained only if the spending by investors at point 2 exactly balances the saving done by consumers at point 1. In other words:

The economy will reach an equilibrium at full employment only if the amount that consumers wish to save out of full-employment incomes happens to be equal to the amount that investors want to invest. If these two magnitudes are unequal, then full employment will not be an equilibrium for the economy.

Specifically, we can see from the circular flow diagram that, if saving exceeds investment at full employment, the total demand received by firms (point 3) will fall short of total output because the added investment spending will not be enough to replace the leakage to saving. With demand inadequate to support production at full employment, GDP must fall below potential. There will be a recessionary gap. Conversely, if investment exceeds saving when the economy is at full employment, then total demand will exceed potential GDP and production will rise above the full-employment level. There will be an inflationary gap.

UNEMPLOYMENT AND INFLATION AS COORDINATION FAILURES

The idea that unemployment stems from a *lack of coordination* between the decisions of savers and investors may seem abstract. But we encounter coordination failures all the time. The following familiar example may bring the idea down to earth. Picture a crowd watching a football game. Now something exciting happens and the fans rise from their seats. People in the front rows begin standing first, and those seated behind them are forced to stand if they want to see the game. Soon everyone in the stadium is on their feet. But with everyone standing, no one can see any better than when everyone was sitting. And the fans are enduring the further discomfort of being on their feet. (Never mind that stadium seats are uncomfortable!) Everyone in the stadium would be better off if everyone sat down. Sometimes this happens. But the crowd rises to its feet again on every exciting play. There is simply no way to coordinate the individual decisions of tens of thousands of football fans.

Unemployment poses a similar coordination problem. During a deep recession, workers are unemployed and businesses cannot sell their wares. Figuratively speaking, everyone is "standing" and unhappy about it. If only the firms could agree to hire more workers, those newly employed people could afford to buy more of the goods and services the firms want to produce. But, as at the football stadium, there is no central authority to coordinate these millions of decisions.

The coordination failure idea also helps to explain why it is so hard to stop inflation. Virtually everyone prefers stable prices to rising prices. But now think of yourself as the seller of a product. If everyone else in the economy would hold their prices steady, you would happily hold yours steady, too. But, if you believe that others will continue to raise their prices at, say, 5 percent per year, you may find it dangerous not to increase yours apace. Hence, society may get stuck with 5 percent inflation even though everyone agrees that zero inflation is better.

Now this discussion does nothing but restate what we already know in different words.[4] But these words hold the key to understanding why the economy can find itself stuck below full employment (or above it, for that matter), for *the people who do the investing are not the same people who do the saving*. In a modern capitalist economy, investing is done by one group of individuals (primarily corporate executives and home buyers) while saving is done by another group.[5] It is easy to imagine that their plans may not be well-coordinated. If they are not, we have just seen how either unemployment or inflation can occur.

Neither of these problems would arise if the acts of saving and investing were not separated in time or space. Imagine a primitive economy of farmers, each of whom invests only in his own farm. There is no borrowing or lending and no financial system. In this world, any farmer wanting to buy a new plow or tractor (that is, wanting to *invest*) would have to refrain from consuming part of his income (that is, would have to *save*). Therefore, the amount that farmers planned to *save* out of full-employment income would have to be equal to the amount they planned to *invest*. Total spending and production would always have to be equal at full employment.

Almost the same holds true in a centrally planned economy. There the state decides how much will be invested and has a great deal of leverage over how much saving people do. If the planners do their calculations correctly, they can force saving to be equal to investment at full employment. Consequently, busi-

[4] In symbols, our equilibrium condition without government or foreign trade is $Y = C + I$. If we note that Y is also the sum of consumption plus saving, $Y = C + S$, it follows that $C + S = C + I$, or $S = I$, is a restatement of the equilibrium condition.

[5] In a modern economy, it is not only households that save. Businesses save also in the form of retained earnings. Nonetheless, households are the ultimate source of the saving needed to finance investment.

ness fluctuations were not historically major problems for the economies of China and the former Soviet Union. (They had plenty of others!) However, as these two countries liberalized their economies, they found that they had to deal with the inflation and unemployment problems that have long plagued the West.

The analysis in the box on the opposite page raises a tantalizing possibility. If both high unemployment and high inflation arise from *coordination failures*, might the government be able to do something about it? Keynes suggested that it could, and we will examine this idea in detail in later chapters. But the football analogy reminds us that a central authority may not find it easy to solve the coordination problem.

SUMMARY

1. *Investment* is the most volatile component of aggregate demand, largely because it is tied so closely to the state of business confidence and to expectations about the future performance of the economy.
2. Government policy cannot influence business confidence in any reliable way, so policies designed to alter investment spending are aimed at more objective, though possibly less important, determinants of investment. Among these are interest rates, the overall state of aggregate demand, and tax incentives.
3. *Net exports* depend on GDPs and relative prices both here and abroad.
4. The **equilibrium** level of national income on the demand side is the level at which total spending just equals the value of production (GDP). Since total spending is the sum of consumption, investment, government purchases, and net exports, the condition for equilibrium is $Y = C + I + G + (X - IM)$.
5. Income levels below equilibrium are bound to rise because, when spending exceeds output, firms will see their inventory stocks being depleted and will react by stepping up production.
6. Income levels above equilibrium are bound to fall because, when total spending is insufficient to absorb total output, inventories will pile up and firms will react by curtailing production.
7. The determination of the equilibrium level of GDP on the demand side can be portrayed on a convenient **income-expenditure diagram** as the point at which the **expenditure schedule**—defined as the sum of $C + I + G + (X - IM)$—crosses the 45° line. The 45° line is significant because it marks off points at which spending and output are equal—that is, at which $Y = C + I + G + (X - IM)$—and this is the basic condition for equilibrium.
8. An income-expenditure diagram can only be drawn up for a specific price level, however. Thus, the equilibrium GDP so determined depends on the price level.
9. Because higher prices reduce the purchasing power of consumers' wealth and hence reduce their spending, equilibrium real GDP demanded is lower when prices are higher. This downward-sloping relationship is known as the *aggregate demand curve*.
10. Equilibrium GDP can be above or below *potential GDP*, which is defined as the GDP that would be produced if the labor force were fully employed.
11. If equilibrium GDP exceeds potential GDP, the difference is called an **inflationary gap**. If equilibrium GDP falls short of potential GDP, the resulting difference is called a **recessionary gap**.
12. Such gaps can occur because the saving that consumers want to do at full-employment income levels may differ from the investing that investors want to do. This problem of *coordination failure* is not likely to arise in a planned economy or in a primitive economy.

KEY TERMS

Investment
Net exports
Equilibrium
Expenditure schedule
Induced investment
$Y = C + I + G + (X - IM)$
Income-expenditure (or 45° line) diagram
Aggregate demand curve
Full-employment level of GDP (or potential GDP)
Recessionary gap
Inflationary gap
Coordination of saving and investment
Coordination failure

QUESTIONS FOR REVIEW

1. For more than 15 years now, imports have consistently exceeded exports in the U.S. economy. This is often considered a major problem. Does this chapter give you any hints about why? (You may want to discuss this issue with your instructor. We will certainly learn more about it in later chapters.)

2. Why is not any arbitrary level of GDP an equilibrium for the economy? (Do not give a mechanical answer to this question. Explain the economic mechanism involved.)

3. From the following data, construct an expenditure schedule on a piece of graph paper. Then use the income-expenditure (45° line) diagram to determine the equilibrium level of GDP.

Income	Consumption	Investment	Government Purchases	Net Exports
$3,600	$3,220	$240	$120	$40
3,700	3,310	240	120	40
3,800	3,400	240	120	40
3,900	3,490	240	120	40
4,000	3,580	240	120	40

4. From the following data, construct an expenditure schedule on a piece of graph paper. Then use the income-expenditure (45° line) diagram to determine the equilibrium level of GDP. Compare your answer with your answer to Review Question 3.

Income	Consumption	Investment	Government Purchases	Net Exports
$3,600	$3,280	$180	$120	$40
3,700	3,340	210	120	40
3,800	3,400	240	120	40
3,900	3,460	270	120	40
4,000	3,520	300	120	40

5. Suppose that investment spending is always $250, government purchases are $100, net exports are always $50, and consumer spending depends on the price level in the following way:

Price Level	Consumer Spending
90	$740
95	720
100	700
105	680
110	660

On a piece of graph paper, use these data to construct an aggregate demand curve. Why do you think this example supposes that consumption declines as the price level rises?

6. Does the economy this year seem to have an inflationary gap or a recessionary gap? (If you do not know the answer from reading the newspaper, ask your instructor.)

7. Why were there no recessions in the former Soviet Union?

8. (More difficult)[6] Consider an economy in which the consumption function takes the following simple algebraic form:

$$C = 300 + 0.75DI$$

and in which investment (I) is always 900 and net exports are always 100. Government purchases are fixed at 1,300 and taxes are fixed at 1,200. Find the equilibrium level of GDP and compare your answer to Table 25-2 and Figure 25-3. (*Hint:* Remember that in this case disposable income is GDP minus taxes: $DI = Y - T = Y - 1{,}200$.)

9. (More difficult) An economy has a consumption function:

$$C = 100 + 0.8DI$$

The government budget is balanced with government purchases and taxes both fixed at 500. Net exports are 100. Investment is 400. Find equilibrium GDP.

[6]The answer to this question is provided in the appendix to this chapter.

APPENDIX THE SIMPLE ALGEBRA OF INCOME DETERMINATION

The model of demand-side equilibrium that the chapter presented graphically and in tabular form can also be handled with some simple algebra. Written as an equation, the consumption function in our example is:

$$\begin{aligned} C &= 300 + 0.75DI \\ &= 300 + 0.75(Y - T) \end{aligned}$$

since, by definition, $DI = Y - T$. This is simply the equation of a straight line with a slope of 0.75 and an intercept of $300 - 0.75T$. Since $T = 1{,}200$ in our example, the intercept is -600 and the equation can be written more simply as:

$$C = -600 + 0.75Y$$

Investment in the example was assumed to be 900, regardless of the level of income, government purchases were 1,300, and net exports were −100. So the sum $C + I + G + (X - IM)$ is:

$$\begin{aligned} &C + I + G + (X - IM) \\ &= -600 + 0.75Y + 900 + 1{,}300 - 100 \\ &= 1{,}500 + 0.75Y \end{aligned}$$

which describes the expenditure curve in Figure 25-3. Since the equilibrium quantity of GDP demanded is defined by:

$$Y = C + I + G + (X - IM)$$

we can solve for the equilibrium value of Y by substituting $1{,}500 + 0.75Y$ for $C + I + G + (X - IM)$ to get:

$$Y = 1{,}500 + 0.75Y$$

To solve this equation for Y, first subtract $0.75Y$ from both sides to get:

$$0.25Y = 1{,}500$$

Then divide both sides by 0.25 to obtain the answer:

$$Y = 6{,}000$$

This, of course, is precisely the solution we found by graphical and tabular methods in the chapter.

We can easily generalize this algebraic approach to deal with any set of numbers in our equations. Suppose that the consumption function is:

$$C = a + bDI = a + b(Y - T)$$

(In the example, $a = 300$, $T = 1{,}200$, and $b = 0.75$.) Then the equilibrium condition that $Y = C + I + G + (X - IM)$ implies that:

$$\begin{aligned} Y &= a + bDI + I + G + (X - IM) \\ &= a - bT + bY + I + G + (X - IM) \end{aligned}$$

Subtracting bY from both sides leads to:

$$(1 - b)Y = a - bT + I + G + (X - IM)$$

and dividing through by $1 - b$ gives:

$$Y = \frac{a - bT + I + G + (X - IM)}{1 - b}$$

This formula, which is certainly *not* to be memorized, is valid for any numerical values of a, b, T, G, I, and $(X - IM)$ (so long as b is between zero and 1).

QUESTIONS FOR REVIEW

1. Find the equilibrium level of GDP demanded in an economy in which investment is always \$400, net exports are always \$150, the government budget is balanced with purchases and taxes both equal to \$600, and the consumption function is described by the following algebraic equation:

   $$C = 50 + 0.75DI$$

   (*Hint:* Do not forget that $DI = Y - T$.)

2. Do the same for an economy in which investment is \$450, net exports are zero, government purchases and taxes are both \$500, and the consumption function is:

   $$C = 250 + 0.5DI$$

3. In each of the above cases, how much saving is there in equilibrium? (*Hint:* Income not consumed must be saved.) Is saving equal to investment?

4. Imagine an economy in which consumer expenditure is represented by the following equation:

   $$C = 100 + 0.75DI$$

Imagine also that investors want to spend 1,000 at every level of income ($I = 1{,}000$), net exports are zero ($X - IM = 0$), government purchases are 600, and taxes are 400.

a. What is the equilibrium level of income?

b. If the full-employment level of income is 6,000, is there a recessionary or inflationary gap? If so, how much?

c. What will happen to the equilibrium level of income if investors become optimistic about the country's future and raise their investment to 1,200?

d. Is there a recessionary or inflationary gap now? How much?

5. Ivyland has the following consumption function:

$$C = 50 + 0.8DI$$

Firms in Ivyland always invest $350 and net exports are zero, initially. The government budget is balanced with spending and taxes both equal to $250.

a. Find the equilibrium level of GDP.

b. How much is saved? Is saving equal to investment?

c. Now suppose that an export-promotion drive succeeds in raising net exports to $50. Answer (a) and (b) under these new circumstances.

CHAPTER 26

CHANGES ON THE DEMAND SIDE: MULTIPLIER ANALYSIS



A definite ratio, to be called the *Multiplier*, can be established between income and investment.
John Maynard Keynes

In the last chapter, we derived the economy's *aggregate demand curve,* which shows how the equilibrium quantity of real GDP demanded depends on the price level—holding all other factors constant. But often these "other factors" do not remain constant, and, as a consequence, the entire aggregate demand curve shifts. This chapter is the first of several that are devoted to enumerating these "other factors" and explaining how and why they make the aggregate demand curve shift.

The central concept of this short chapter is the *multiplier*—the idea that an increase in spending will bring about an *even larger* increase in equilibrium GDP. We approach this idea from three different perspectives, each of which provides different insights into the multiplier process. First, we illustrate the multiplier graphically using the income-expenditure diagram from Chapter 25. Next, we reach the same conclusion through the use of a numerical example, and finally, we offer an algebraic statement. Each of these is an expression of the remarkable multiplier result. Near the end of the chapter, we use multiplier analysis to explain how economic developments abroad affect the U.S. economy and why a drive to increase national saving might not succeed.

CASE STUDY: HOW DID KANSAS DO IT?

In 1989, the Kansas state legislature voted to spend $2.6 billion over the next several years to make long overdue repairs on some 135,000 miles of highway. Everyone knew that spending on that scale would give a big boost to the state's construction industry. But Kansans were in for a pleasant surprise. Within 3 years, all sorts of industries, including many with no obvious connection to road building, were booming in Kansas. Personal income in the state was growing at twice the national average. And the state's unemployment rate was the fourth-lowest in the nation.

Did spending on road repair contribute to the boom in other industries? If so, how? This chapter will provide some answers.

THE MAGIC OF THE MULTIPLIER

Because it is subject to such abrupt swings, investment spending is often the cause of business fluctuations in the United States and elsewhere. So let us ask what would happen to equilibrium income in our fictitious country, Macroland, if firms there suddenly decided to spend more on investment goods. As we shall see, such a decision would have a *multiplied* effect on GDP in Macroland; that is, each $1 of additional investment spending would add more than $1 to GDP. The same would be true in Kansas, or in the entire U.S. economy.

For simplicity, we continue to assume that the price level is fixed—an assumption we will drop in the next chapter. Refer first to Table 26-1, which looks very much like Table 25-1 (page 595). The only difference is that we assume here that, for some reason, firms in Macroland now want to invest $200 billion more than they previously did—for a total of $1,100 billion.

The **MULTIPLIER** is the ratio of the change in equilibrium GDP (Y) divided by the original change in spending that causes the change in GDP.

The **multiplier** principle says that Macroland's GDP will rise by more than the $200 billion increase in investment. Specifically, the multiplier is defined as the ratio of the change in equilibrium GDP (Y) to the original change in spending that causes GDP to change. In shorthand, when we deal with the multiplier for investment (I), the formula is:

$$\text{Multiplier} = \frac{\text{Change in } Y}{\text{Change in } I}$$

Let us verify that the multiplier is indeed greater than 1. Table 26-1 shows how to derive a new expenditure schedule by adding up C, I, G, and $(X - IM)$ at each level of Y, just as we did in Chapter 25. If you compare the last column of Table 26-1 with that of Table 25-1, you will see that the new expenditure schedule lies uniformly above the old one by $200 billion.

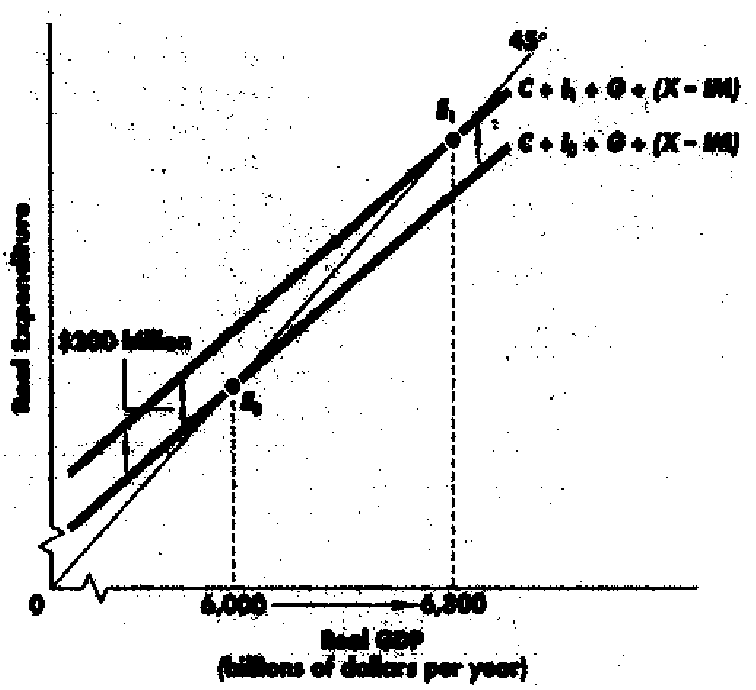
Figure 26-1 illustrates this shift graphically. The schedule marked $C + I_0 + G + (X - IM)$ is derived from the last column of Table 25-1, while the higher schedule marked $C + I_1 + G + (X - IM)$ is derived from the last column of Table 26-1. The two expenditure lines are parallel and $200 billion apart.

TABLE 26-1 TOTAL EXPENDITURE AFTER A $200 BILLION RISE IN INVESTMENT SPENDING (BILLIONS OF DOLLARS PER YEAR)

(1) Income (Y)	(2) Consumption (C)	(3) Investment (I)	(4) Government Purchases (G)	(5) Net Exports (X − IM)	(6) Total Expenditure
4,800	3,000	1,100	1,300	−100	5,300
5,200	3,300	1,100	1,300	−100	5,600
5,600	3,600	1,100	1,300	−100	5,900
6,000	3,900	1,100	1,300	−100	6,200
6,400	4,200	1,100	1,300	−100	6,500
6,800	**4,500**	**1,100**	**1,300**	**−100**	**6,800**
7,200	4,800	1,100	1,300	−100	7,100

This table shows the construction of a total expenditure schedule for Macroland after investment has risen to $1,100 billion. As indicated by the shaded numbers, only income level Y = $6,800 billion is an equilibrium on the demand side of the economy because only at this level is total spending, $C + I + G + (X - IM)$, equal to production (Y).

FIGURE 26-1 ILLUSTRATION OF THE MULTIPLIER



This figure depicts the multiplier effect of a rise in investment spending of $200 billion. The expenditure schedule shifts upward from $C + I_0 + G + (X - IM)$ to $C + I_1 + G + (X - IM)$, thus moving equilibrium from point E_0 to point E_1. The rise in income is $800 billion, so the multiplier is $800/$200 = 4.

So far no act of magic has occurred—things look just as you might expect. But one more step will bring the multiplier rabbit out of the hat. Let us see what the upward shift of the expenditure line does to equilibrium income. In Figure 26-1, equilibrium moves outward from point E_0 to point E_1; that is, from $6,000 billion to $6,800 billion. The difference is an increase in national income of $800 billion. All this from a $200 billion stimulus to investment? That is the magic of the multiplier.

Because the change in I is $200 billion and the change in equilibrium Y is $800 billion, by applying our definition, the multiplier is:

$$\text{Multiplier} = \frac{\text{Change in } Y}{\text{Change in } I} = \frac{\$800}{\$200} = 4$$

This tells us that, in our example, every additional dollar of investment demand will add $4 to equilibrium GDP!

This does indeed seem mysterious. Can something be created from nothing? Let us first check that the graph has not deceived us. The first and last columns of Table 26-1 show in numbers what Figure 26-1 shows in a picture. Notice that equilibrium now comes at Y = $6,800 billion. This equilibrium level of GDP is $800 billion higher than the $6,000 billion level found in the last chapter, when investment was only $900 billion. Thus, a $200 billion rise in investment leads to an $800 billion rise in equilibrium GDP. The multiplier really is 4.

DEMYSTIFYING THE MULTIPLIER: HOW IT WORKS

The multiplier result seems implausible at first, but it loses its mystery once we remember the circular flow of income and expenditure, and the simple fact that one person's spending is another person's income. To illustrate the logic of the

multiplier and see why it is exactly 4 in our model economy, let us look more closely at what actually happens if businesses decide to spend an additional $1 million on investment goods.

Suppose that Generous Motors—a major corporation in Macroland—decides to spend $1 million to retool a factory to manufacture cars powered by compressed natural gas. Its $1 million expenditure goes to construction workers and owners of construction companies as wages and profits. That is, it becomes their *income.*

But the owners and workers of the construction firms will not keep their $1 million in the bank. They will spend most of it. If they are "typical" consumers, their spending will be $1 million times the marginal propensity to consume (MPC). In our example, the MPC is 0.75. So let us assume that they spend $750,000 and save the rest. *This $750,000 expenditure is a net addition to the nation's demand for goods and services exactly as Generous Motors' original $1 million expenditure was.* So, at this stage, the $1 million investment has already pushed GDP up some $1.75 million. But the process by no means stops here.

Shopkeepers receive the $750,000 spent by construction workers, and they in turn also spend 75 percent of their new income. This accounts for $562,500 (75 percent of $750,000) in additional consumer spending in the "third round." Next follows a fourth round in which the recipients of the $562,500 spend 75 percent of this amount, or $421,875, and so on. At each stage in the spending chain, people spend 75 percent of the additional income they receive, and the process continues. Consumption grows in each round.

Where does it all end? Does it all end? The answer is that it does, indeed, eventually end—with GDP a total of $4 million higher than it was before Generous Motors spent the original $1 million. The multiplier, is, indeed, 4.

Table 26-2 displays the basis for this conclusion. In the table, "round 1" represents Generous Motors' initial investment, which creates $1 million in income for construction workers; "round 2" represents the construction workers' spending which creates $750,000 in income for shopkeepers. The rest of the table proceeds accordingly; each entry in Column 2 is 75 percent of the previous entry. Column 3 tabulates the running sum of Column 2.

We see that after 10 rounds of spending, the initial $1 million investment has mushroomed to $3.77 million, and the sum is still growing. After 20 rounds, the total increase in GDP is over $3.98 million—near its eventual value of $4 million. While it takes quite a few rounds of spending before the multiplier chain is near 4, we see from the table that it hits 3 rather quickly. If each income recipient in the chain waits, say, 2 months before spending his new income, the multiplier will reach 3 in only about 10 months.

Figure 26-2 provides a graphical presentation of the numbers in the last column of Table 26-2. Notice how the multiplier builds up rapidly at first and then tapers off to approach its ultimate value (4 in this example) gradually.

While this is only a hypothetical example, the same thing occurs every day in the real world. For instance, a burst of new housing starts has a multiplier effect on everything from appliances and furniture to carpeting and insulation. Similarly, when a large company like AT&T makes an investment in a developing country, the multiplier effect boosts business activity in many sectors.

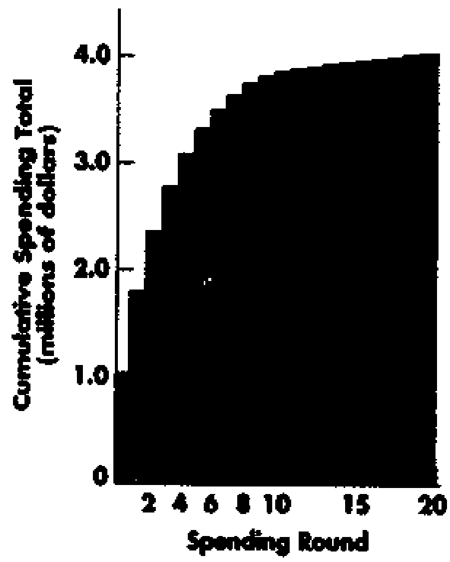
TABLE 26-2 THE MULTIPLIER SPENDING CHAIN

This table shows how the multiplier unfolds through time. Round 1 is Generous Motors' initial spending, which leads to $1 million in additional income to construction workers. Round 2 shows the construction workers spending 75 percent of this amount, since the marginal propensity to consume is 0.75. The other rounds proceed accordingly, with spending in each successive round equal to 75 percent of that in the previous round. Technically, the full multiplier of 4 is reached only after an "infinite" number of rounds. But, as can be seen, we are very close to the full amount after 20 rounds.

(1) Round Number	(2) Spending in This Round	(3) Cumulative Total
1	$1,000,000	$1,000,000
2	750,000	1,750,000
3	562,500	2,312,500
4	421,875	2,734,375
5	316,406	3,050,781
6	237,305	3,288,086
7	177,979	3,466,065
8	133,484	3,599,549
9	100,113	3,699,662
10	75,085	3,774,747
⋮	⋮	⋮
20	4,228	3,987,317
⋮	⋮	⋮
"Infinity"	0	4,000,000

FIGURE 26-2 HOW THE MULTIPLIER BUILDS

This diagram portrays the numbers from Table 26-2 and shows how the multiplier builds through time. Notice how the effect grows quickly at first and how the full effect is almost reached after 10 rounds.



ALGEBRAIC STATEMENT OF THE MULTIPLIER

Figure 26-2 and Table 26-2 probably make a persuasive case that the multiplier eventually reaches 4. But for the remaining skeptics we offer a simple algebraic proof.[1] Most of you learned about something called an *infinite geometric progression* in high school. This term refers to an infinite series of numbers, each one of which is a fixed fraction of the previous one. The fraction is called the *common ratio*. A geometric progression beginning with 1 and having a common ratio of 0.75 would look like this:

$$1 + 0.75 + (0.75)^2 + (0.75)^3 + \ldots$$

More generally, a geometric progression beginning with 1 and having a common ratio R would be:

$$1 + R + R^2 + R^3 \ldots$$

A simple formula enables us to sum such a progression as long as R is less than 1.[2] The formula is:[3]

$$\text{Sum of infinite geometric progression} = \frac{1}{1 - R}$$

Now we can recognize that the multiplier chain in Table 26-2 is just an infinite geometric progression with 0.75 as its common ratio. That is, each \$1 spent by Generous Motors leads to a $(0.75) \times \$1$ expenditure by construction workers, which in turn leads to a $(0.75) \times (0.75 \times \$1) = (0.75)^2 \times \$1$ expenditure by the shopkeepers, and so on. Thus, for each initial dollar of investment spending, the progression is:

$$1 + 0.75 + (0.75)^2 + (0.75)^3 + (0.75)^4 + \ldots$$

Applying the formula for the sum of such a series, we find that:

$$\text{Multiplier} = \frac{1}{1 - 0.75} = \frac{1}{0.25} = 4$$

Notice how this result can be generalized. If we did not have a specific number for the marginal propensity to consume, but simply called it MPC, the geometric progression in Table 26-2 would have been:

$$1 + MPC + (MPC)^2 + (MPC)^3 + \ldots$$

---

[1]Students who blanch at the sight of algebra should not be put off. Anyone who can balance a checkbook (even many who cannot!) will be able to follow the argument.

[2]If R exceeds 1, nobody can possibly sum it—not even with the aid of a modern computer—because the sum is not a finite number.

[3]The proof of the formula is simple. Let the symbol S stand for the (unknown) sum of the series:

$$S = 1 + R + R^2 + R^3 + \ldots$$

Then, multiplying by R,

$$RS = R + R^2 + R^3 + R^4 + \ldots$$

By subtracting RS from S, we obtain:

$$S - RS = 1 \quad \text{or} \quad S = \frac{1}{1 - R}$$

which has the MPC as its common ratio. Applying the same formula for summing a geometric progression to this more general case gives us the following general result:

OVERSIMPLIFIED FORMULA FOR THE MULTIPLIER

$$\text{Multiplier} = \frac{1}{1 - \text{MPC}}$$

We call this formula "oversimplified" because it ignores many factors that are important in the real world. One of them is *international trade*—in particular, the fact that a country's imports depend on its GDP. We deal with this complication in Appendix B. A second factor is *inflation*, a complication we will address in the next chapter. A third is *income taxation*, a point we will elaborate in Chapter 28. The last important influence arises from the *financial system* and, after we discuss money and banking in Chapters 29 and 30, we will explain it in Chapter 31. As it turns out, each of these factors *reduces* the size of the multiplier.

We can begin to appreciate just how unrealistic the "oversimplified" formula is by considering some real numbers for the U.S. economy. The marginal propensity to consume (MPC) has been estimated many times and is about 0.9. From our oversimplified formula, then, it would seem that the multiplier should be:

$$\text{Multiplier} = \frac{1}{1 - 0.9} = \frac{1}{0.1} = 10$$

An **INDUCED INCREASE IN CONSUMPTION** is an increase in consumer spending that stems from an increase in consumer incomes. It is represented on a graph as a movement along a fixed consumption function.

In fact, the actual multiplier for the U.S. economy is believed to be less than 2. This is quite a discrepancy! But it does not mean that anything we have said about the multiplier so far is incorrect. Our story is simply incomplete. As we progress through this and subsequent chapters, you will learn why the multiplier is below 2 even though the MPC is close to 0.9. For now, we simply point out that:

While the multiplier is larger than 1 in the real world, it cannot be calculated with any degree of accuracy from the oversimplified formula. The actual multiplier is *much lower* than the formula suggests.

THE MULTIPLIER EFFECT OF CONSUMER SPENDING

Business firms that invest are not the only ones that can work the magic of the multiplier; so can consumers. To see how the multiplier works when the process is initiated by an upsurge in consumer spending, we must distinguish between two types of change in consumer spending.

An **AUTONOMOUS INCREASE IN CONSUMPTION** is an increase in consumer spending without any increase in incomes. It is represented on a graph as a shift of the entire consumption function.

When C rises because income rises—that is, when consumers move outward *along a fixed consumption function*—we call the increase in C an **induced increase in consumption.** However, if instead C rises because the entire consumption function *shifts up*, we call this an **autonomous increase in consumption.** The name indicates that consumption changes independently of income, and Chapter 24's discussion pointed out that a number of events, such as a change in the price level or in the value of the stock market, can initiate such a shift.

Let us suppose that, for some reason, consumer spending rises autonomously by $200 billion. In this case, we would revise our table of aggregate demand to

TABLE 26-3 TOTAL EXPENDITURE AFTER CONSUMERS DECIDE TO SPEND $200 BILLION MORE (BILLIONS OF DOLLARS PER YEAR)

(1) Income (*Y*)	(2) Consumption (*C*)	(3) Investment (*I*)	(4) Government Purchases (*G*)	(5) Net Exports (*X* − *IM*)	(6) Total Expenditure
4,800	3,200	900	1,300	−100	5,300
5,200	3,500	900	1,300	−100	5,600
5,600	3,800	900	1,300	−100	5,900
6,000	4,100	900	1,300	−100	6,200
6,400	4,400	900	1,300	−100	6,500
6,800	**4,700**	**900**	**1,300**	**−100**	**6,800**
7,200	5,000	900	1,300	−100	7,100

This table shows the construction of the total expenditure schedule for Macroland following an autonomous increase of $200 billion in consumption rather than in investment. Notice that Columns 2 and 3 differ from the corresponding columns in Table 26-1, but Column 6 is the same in both tables. Thus, the expenditure schedule in the 45° line diagram is the same as in the earlier example.

look like Table 26-3. Comparing this to Table 26-1 on page 610, we note that each entry in Column 2 is $200 billion *higher* than the corresponding entry in Table 26-1 (because consumption is higher), and each entry in Column 3 is $200 billion *lower* (because investment is lower).

The equilibrium level of income is clearly Y = $6,800 billion once again. Indeed, the entire expenditure schedule (Column 6) is the same as it was in Table 26-1. The initial rise of $200 billion in spending leads to an ultimate rise of $800 billion in GDP, just as occurred in the case of higher investment spending. In fact, Figure 26-1 applies directly to this case once we note that the upward shift is now caused by an autonomous change in *C* rather than in *I*. The multiplier for autonomous changes in consumer spending, then, is also 4 (= $800/$200).

The reason is straightforward. It does not matter who injects an additional dollar of spending into the economy, whether it is business investors or consumers. Wherever it comes from, 75 percent of it will be respent if the MPC is 0.75, and the recipients of this second round will, in turn, spend 75 percent of their additional income, and so on and on. And that is what constitutes the multiplier process.

THE MULTIPLIER EFFECT OF GOVERNMENT PURCHASES

What about the third component of total spending, government purchases (*G*)? Since we now know that the multiplier process is the same no matter who injects the additional dollar of spending into the economy, we conclude that *G* has to have the very same multiplier as *I* and *C*. Figure 26-1 can again be used to illustrate the conclusion graphically—just think of the upward shift as being caused by a change in *G* this time.

The multipliers are identical because the logic behind them is identical. The multiplier spending chain set in motion when Generous Motors spent $1 million to build a factory could equally well have been kicked off by the government buying $1 million worth of new cars from Generous Motors. Thereafter, each recipient of additional income would spend 75 percent of it (the assumed mar-

ginal propensity to consume), until $4 million in new income had eventually been created.

The idea that changes in *G* have multiplier effects on GDP will play a central role in the discussion of government stabilization policy that begins in Chapter 28. So it is worth noting here that:

Changes in the volume of government purchases of goods and services will change the equilibrium level of GDP in the same direction, and by a multiplied amount.

This is, more or less, what happened in Kansas starting in 1989. In the "first round," the state government spent money on road building. But in the "second round," the construction workers spent most of their paychecks on unrelated (to road building) items like groceries, paint, and toys for their kids. In the "third" and subsequent rounds, yet more recipients of new income purchased still different goods and services. Pretty soon, the whole state economy was booming.

THE MULTIPLIER EFFECT OF NET EXPORTS

At this point, it will not surprise you to learn that a change in net exports has precisely the same multiplier effect on equilibrium GDP as a change in any of the other components of spending. The reason is hardly mysterious. When foreigners buy U.S. products, they put income into the hands of Americans, just as domestic investment does. As this income is spent and respent, a multiplier process is set in motion, raising GDP.

Although we will have much more to learn about how the U.S. economy is linked to the economies of other countries in Part VIII, this simple analysis of the multiplier effect of foreign trade already teaches us an important lesson: *Booms and recessions tend to be transmitted across national borders.*

Why is that? Suppose a boom abroad raises aggregate demand and GDP in foreign countries. With rising incomes, foreigners will buy more American goods—which means that U.S. exports will rise. But a rise in our exports will, via the multiplier, raise GDP in the United States. By this mechanism, rapid economic growth abroad contributes to rapid economic growth here.

Of course, the same mechanism also operates in the downward direction. Suppose some of the countries that trade with us slip into recession. As their GDPs decline, so do their *imports*. But this means that the United States will experience a decline in *exports* which, through the multiplier, will pull down GDP here. Hence, a recession abroad can contribute to recessionary conditions in the United States.

Naturally, what foreign countries do to us, we also do to them. Thus, rapid economic growth in the United States tends to produce boom conditions in the countries from which we buy, and recessions here tend quickly to spill beyond our borders. In summary:

The GDPs of the major economies are linked by trade. A boom in one country tends to raise its imports and hence push up exports and GDP in other countries. Similarly, a recession in one country tends to pull GDP down in other countries.

So, for example, strong economic growth in the United States in 1993 and 1994 helped pull several European countries out of recession, just as recession in the United States in 1990–1991 had hurt Europe.

THE MULTIPLIER IN REVERSE

A good way to check your understanding of the multiplier process is to run it in reverse: What happens if, for example, consumers abruptly decide to spend less? For example, suppose a wave of thriftiness comes over the people of Macroland so that, no matter what their total income, they now want to spend $200 billion *less* than they did previously rather than the $200 billion *more* assumed in Table 26-1.

A decision to spend $200 billion less out of any given level of income is, by definition, a *downward* shift of the total expenditure schedule by $200 billion. This is shown in Figure 26-3, where the $C + I + G + (X - IM)$ schedule falls from $C_0 + I + G + (X - IM)$ to $C_1 + I + G + (X - IM)$. The horizontal distance between these two parallel lines is the $200 billion drop in spending.

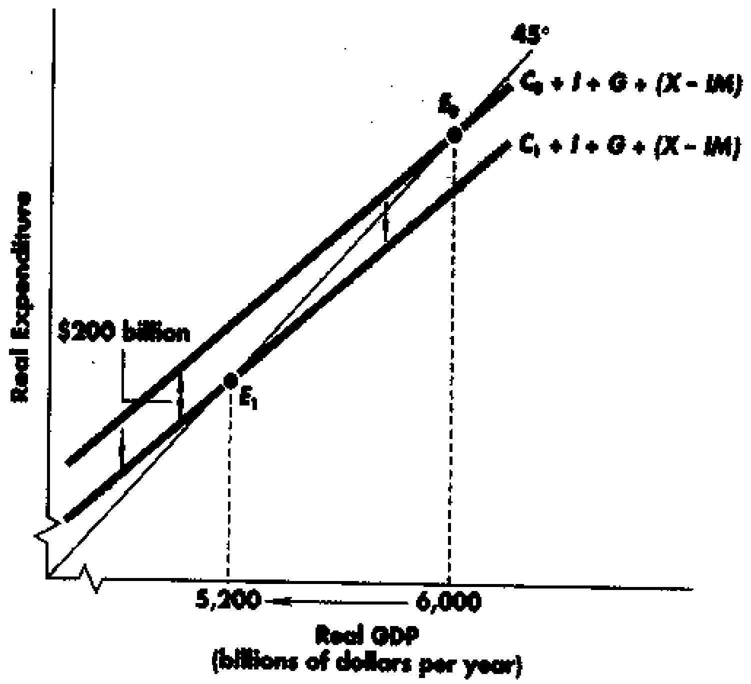
There are two ways to calculate the multiplier. First, our oversimplified multiplier formula tells us that the multiplier is:

$$\frac{1}{1 - \text{MPC}} = \frac{1}{1 - 0.75} = \frac{1}{0.25} = 4$$

So a $200 billion drop in spending will lead to a multiplier effect of $800 billion. Alternatively, we can read this conclusion from Figure 26-3. Here the economy's equilibrium point moves down the 45° line from point E_0 to E_1; income drops from $6,000 billion to $5,200 billion—a decline of $800 billion.

Now compare the analysis of a decline in spending summarized in Figure 26-3 with the previous analysis of an increase in spending shown in Figure 26-1 on page 611. You will see that everything is simply turned in the opposite direction. The multiplier works in both directions.

FIGURE 26-3 THE MULTIPLIER IN REVERSE



This diagram shows the multiplier effect of an autonomous decline in consumer spending of $200 billion. The decline appears as a downward shift of $200 billion in the expenditure schedule, which falls from $C_0 + I + G + (X - IM)$ to $C_1 + I + G + (X - IM)$. Equilibrium, which is always at the intersection of the expenditure schedule and the 45° line, moves from point E_0 to point E_1, and income falls from $6,000 billion to $5,200 billion.

THE PARADOX OF THRIFT

This last example of multiplier analysis teaches us an important lesson: an increase in the desire to save will lead to a cumulative fall in GDP. And, *because saving depends on income,* the resulting decline in national income will pull saving down. In consequence, saving may fail to rise despite an increase in the public's desire to save.

Let us be a bit more specific about this remarkable result. Before the upsurge in saving, consumers were spending $3,900 billion out of a total national income of $6,000 billion, as we can see in Table 25-1 on page 595. How much was being saved? Since taxes in Macroland are assumed to be fixed at $1,200 billion, disposable income was:

$$DI = Y - T = \$6{,}000 - \$1{,}200 = \$4{,}800 \text{ billion}$$

Hence, saving was $900 billion (= $4,800 − $3,900).

In Figure 26-3, GDP falls to $5,200 billion, so disposable income drops to $4,000. Since investment, government purchases, and net exports are all unchanged, the entire $800 billion drop in GDP must come out of consumption, which therefore falls by $800 billion (to $3,100 billion). Thus, DI is down to $4,000 billion and C is down to $3,100 billion, leaving total saving still $900 billion. The effort to save more has been totally frustrated by the decline in GDP.[4]

The **PARADOX OF THRIFT** is the fact that an effort by a nation to save more may simply reduce national income and fail to raise total saving.

This remarkable result is called the **paradox of thrift,** because it shows that, while saving may pave the road to riches for an individual, if the nation as a whole decides to save more, the result may be a recession and the falling incomes that come with it. The paradox of thrift is important because it is contrary to most people's thinking, and it means that a greater desire to save may be a mixed blessing if it is not accompanied by greater desire to invest.

THE MULTIPLIER AND THE AGGREGATE DEMAND CURVE

At this point, we must recall something mentioned at the start of the chapter: Income-expenditure diagrams such as Figures 26-1 and 26-3 can be drawn up only for a given price level. A different price level leads to a different total expenditure curve. This means that our oversimplified multiplier formula measures *the increase in real GDP demanded that would occur if the price level were fixed.* That is, it measures the *horizontal shift* of the economy's aggregate demand curve.

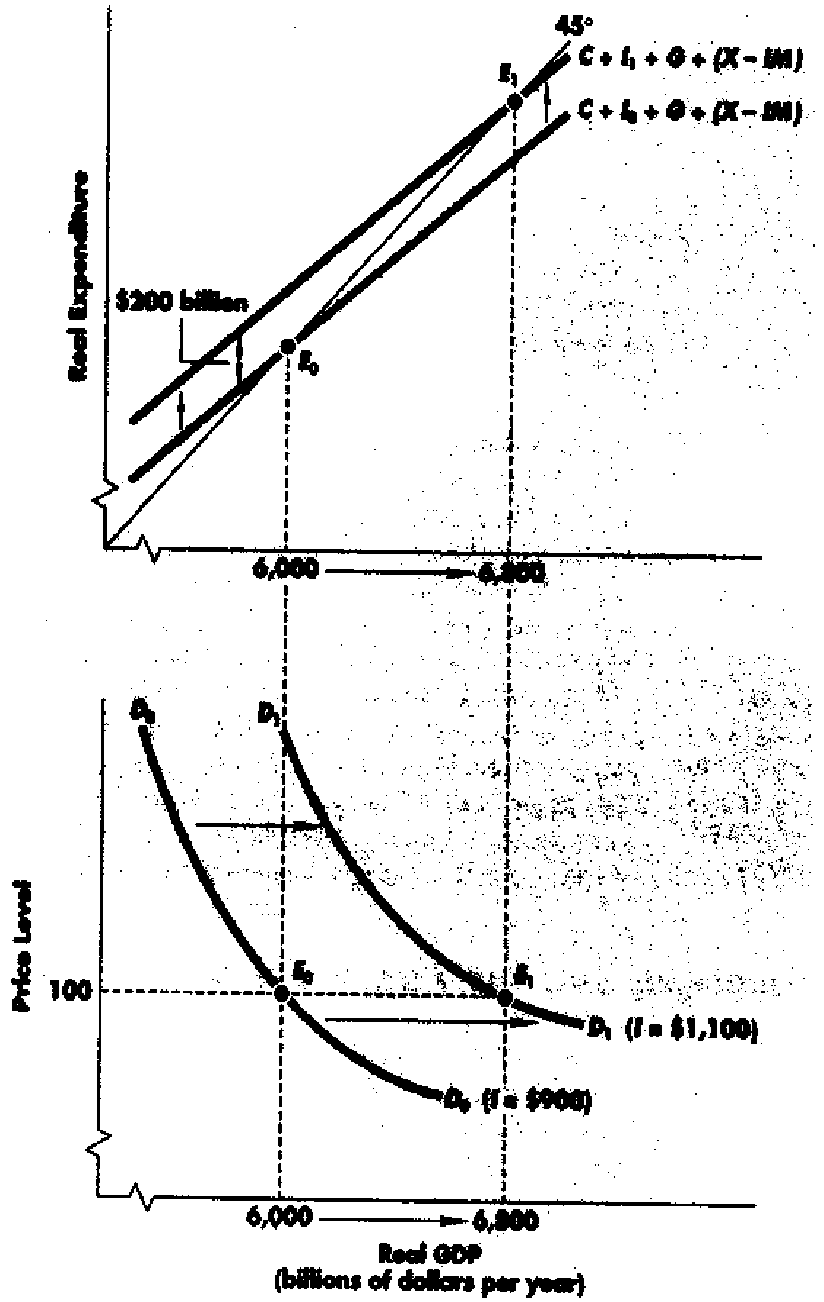
Figure 26-4 illustrates this conclusion by supposing that the price level that underlies Figure 26-1 is $P = 100$. The top panel simply repeats Figure 26-1 and shows how an increase in investment spending from $900 to $1,100 billion leads to an increase in GDP from $6,000 to $6,800 billion.

The bottom panel shows two downward-sloping aggregate demand curves. The first, labeled D_0D_0, depicts the situation when investment is $900 billion. Point E_0 on this curve indicates that, at the given price level ($P = 100$), the equilibrium quantity of GDP demanded is $6,000 billion. It corresponds exactly to point E_0 in the top panel. The second aggregate demand curve, D_1D_1, depicts the situation after investment has risen to $1,100 billion. Point E_1 on this curve

[4] It is even possible to devise examples in which total saving goes *down* when people attempt to save more. This will happen, for example, if there is *induced investment.*

FIGURE 26-4 TWO VIEWS OF THE MULTIPLIER

The top panel repeats Figure 26-1. The bottom panel shows two aggregate demand curves. Curve D_0D_0, which applies when investment is $900 billion, shows that equilibrium GDP on the demand side comes at Y = $6,000 billion when P = 100 (point E_0). Curve D_1D_1, which applies when investment is $1,100 billion, shows that equilibrium GDP on the demand side comes at Y = $6,800 billion when P = 100 (point E_1). The horizontal distance between points E_0 and E_1 in the bottom panel indicates the oversimplified multiplier effect.



indicates that the equilibrium quantity of GDP demanded when P = 100 has risen to $6,800 billion, which corresponds exactly to point E_1 in the top panel.

As Figure 26-4 shows, the horizontal distance between the two aggregate demand curves is exactly equal to the increase in real GDP shown in the income-expenditure diagram—in this case, $800 billion. Thus:

An autonomous increase in spending leads to a horizontal shift of the aggregate demand curve by an amount given by the oversimplified multiplier formula.

Thus, everything we have learned about the multiplier applies to *shifts of the economy's aggregate demand curve*. If businesses decide to increase their investment

spending, or if the consumption function shifts up, or if the government or foreigners decide to buy more goods, the aggregate demand curve moves horizontally to the right—as indicated in Figure 26-4. If any of these variables move down instead, the aggregate demand curve moves horizontally to the left.

Thus, the economy's aggregate demand curve cannot be expected to stand still for long. Autonomous changes in one or another of the four components of total spending will cause the aggregate demand curve to move around. But to understand the consequences of shifts of aggregate demand, we must bring the aggregate supply curve into the picture. That is the task of the next chapter.

SUMMARY

1. Any autonomous increase in expenditure has a **multiplier** effect on GDP; that is, it increases GDP by more than the original increase in spending.
2. The reason for this multiplier effect is that one person's additional expenditure constitutes a new source of income for another person, and this additional income leads to still more spending, and so on.
3. The multiplier also works in reverse: An autonomous decrease in any component of aggregate demand leads to a multiplied decrease in national income.
4. The multiplier is the same for an **autonomous increase in consumption**, investment, government purchases, or net exports.
5. A simple formula for the multiplier says that its numerical value is 1/(1 − MPC). This formula, which is too simple to give accurate results, measures the horizontal shift of the aggregate demand curve.
6. Rapid (or sluggish) economic growth in one country contributes to rapid (or sluggish) growth in other countries because one country's imports are other countries' exports.
7. If the nation as a whole decides to save more, that is, to consume less, the resulting decline in national income may serve to make everyone poorer. This possibility that thriftiness, while a virtue for the individual, may be disastrous for an entire nation, is called the **paradox of thrift**.

KEY TERMS

The multiplier
Induced increase in consumption
Autonomous increase in consumption
Paradox of thrift

QUESTIONS FOR REVIEW

1. Try to remember where you last spent a dollar. Explain how this dollar will lead to a multiplier chain of increased income and spending. (Who received the dollar? What will he or she do with it?)
2. Use both numerical and graphical methods to find the multiplier effect of the following shift in the consumption function in an economy in which investment is always $220, government purchases are always $100, and net exports are always $40. (*Hint:* What is the marginal propensity to consume?)

Income	Consumption before Shift	Consumption after Shift
1,080	880	920
1,140	920	960
1,200	960	1,000
1,260	1,000	1,040
1,320	1,040	1,080
1,380	1,080	1,120
1,440	1,120	1,160
1,500	1,160	1,200

3. Turn back to Review Question 3 in Chapter 25 (page 606). Suppose investment spending rises to \$260, and the price level is fixed. By how much will the equilibrium GDP increase? Derive the answer both numerically and graphically.
4. Explain the paradox of thrift. Why do you think it is called a paradox?
5. (More difficult) Suppose the consumption function is as given in Review Question 8 of Chapter 25 (page 606):

$$C = 300 + 0.75DI$$

and investment (I) rises to 1,100 while net exports ($X - IM$) remain at -100, government purchases remain at 1,300, and taxes remain at 1,200. Use the equilibrium condition $Y = C + I + G + (X - IM)$ to find the equilibrium level of GDP. (In working out the answer, assume the price level is fixed.) Compare your answer to Table 26-1 and Figure 26-1. Now compare your answer to the answer to Review Question 8 of Chapter 25. What do you learn about the multiplier?
6. (More difficult) Look back at Review Question 9 of Chapter 25 (page 606). What is the multiplier for this economy? If G rises by 100, what happens to Y? What happens to Y if both G and T rise by 100 at the same time?

APPENDIX A THE SIMPLE ALGEBRA OF THE MULTIPLIER

The appendix to Chapter 25 presented a general expression for the equilibrium level of GDP when the price level is fixed, investment (I), government purchases (G), taxes (T), and net exports ($X - IM$) are all constant, and the consumption function is:

$$C = a + bDI = a + b(Y - T)$$

The answer obtained there (which can be found on page 607) was:

$$Y = \frac{a - bT + I + G + (X - IM)}{1 - b}$$

From this formula, it is easy to derive the oversimplified multiplier formula algebraically and to show that it applies equally well to a change in investment, autonomous consumer spending, government purchases, or net exports. To do so, suppose that any of the symbols in the numerator of the multiplier formula increases by 1 unit. In any of these cases, GDP would rise from the previous formula to:

$$Y = \frac{a - bT + I + G + (X - IM) + 1}{1 - b}$$

By comparing this with the previous expression for Y, we see that a 1-unit change in any component of spending changes equilibrium GDP by:

$$\text{Change in } Y = \frac{a - bT + I + G + (X - IM) + 1}{1 - b} - \frac{a - bT + I + G + (X - IM)}{1 - b}$$

or

$$\text{Change in } Y = \frac{1}{1 - b}$$

Recalling that b is the marginal propensity to consume, we see that this is precisely the oversimplified multiplier formula.

APPENDIX B THE MULTIPLIER WITH VARIABLE IMPORTS

In Chapters 25 and 26, we assumed that net exports were a fixed number. But in fact a nation's imports depend on its GDP. The reason is simple; higher GDP leads to higher incomes, some of which is spent on foreign goods. Thus:

Our imports rise as our GDP rises and fall as our GDP falls.

Similarly, our *exports* are the *imports* of other countries, so it is natural to assume that our

TABLE 26-4 EQUILIBRIUM INCOME WITH VARIABLE IMPORTS (BILLIONS OF DOLLARS PER YEAR)

(1) Gross Domestic Product (Y)	(2) Consumer Expenditures (C)	(3) Investment (I)	(4) Government Purchases (G)	(5) Exports (X)	(6) Imports (IM)	(7) Net Exports (X − IM)	(8) Total Expenditure [C + I + G + (X − IM)]
4,800	3,000	900	1,300	650	570	+80	5,280
5,200	3,300	900	1,300	650	630	+20	5,520
5,600	3,600	900	1,300	650	690	−40	5,760
6,000	**3,900**	**900**	**1,300**	**650**	**750**	**−100**	**6,000**
6,400	4,200	900	1,300	650	810	−160	6,240
6,800	4,500	900	1,300	650	870	−220	6,480
7,200	4,800	900	1,300	650	930	−280	6,720

exports depend on *their* GDPs, not on our own. Thus:

Our exports are relatively insensitive to our own GDP, but are quite sensitive to the GDPs of other countries.

This appendix derives the implications of these rather elementary observations. In particular, it shows that once we recognize the dependence of a nation's imports on its GDP:

International trade lowers the value of the multiplier.

To see why, we begin with Table 26-4, which adapts the example of Macroland to allow imports to depend on GDP. Columns 2 through 4 are the same as in Table 25-1 on page 595; they show *C*, *I*, and G at alternative levels of GDP. Columns 5 and 6 record revised assumptions about the behavior of exports and imports. Exports are fixed at $650 billion regardless of GDP. But imports are assumed to rise by $60 billion for every $400 billion rise in GDP, which is a simple numerical example of the idea that imports depend on GDP. Column 7 subtracts imports from exports to get net exports, $(X - IM)$, and Column 8 adds up the four components of total expenditure, $C + I + G + (X - IM)$. The equilibrium, you can see, occurs at Y = $6,000 billion, just as it did in Chapter 25.

Figures 26-5 and 26-6 display the same conclusion graphically. The upper panel of Figure 26-5 shows that exports are fixed at $650 billion regardless of GDP while imports increase as GDP rises, just as in Table 26-4. The difference between exports and imports, or net exports, is positive until GDP reaches around $5,300 billion and negative once GDP surpasses that amount. The bottom panel of Figure 26-5 shows the subtraction explicitly and makes it clear that:

Net exports decline as GDP rises.

Figure 26-6 carries this analysis over to the 45° line diagram. We begin with the familiar $C + I + G + (X - IM)$ line of Chapters 25 and 26 in black. There we simply assumed that net exports were fixed at −$100 billion regardless of GDP. Now that we have amended our model to note that net exports decline as GDP rises, the sum $C + I + G + (X - IM)$ rises more slowly than we previously assumed. This is shown by the orange line. Note that it is less steep than the black line.

Let us now consider what happens if exports rise by $160 billion while imports remain as in Table 26-4. Table 26-5 shows that equilibrium now occurs at a GDP of Y = $6,400 billion. Naturally, higher exports have raised domestic GDP. But consider the magnitude. A $160 billion increase in exports (from $650 billion to $810 billion) leads to an increase of $400 billion in GDP (from $6,000 billion to $6,400 billion). So the multiplier is 2.5 (= $400/$160).[5]

---

[5]EXERCISE: Construct a version of Table 26-4 to show what would happen if imports rose by $160 billion at every level of GDP while exports remained at $650 billion. You should be able to show that the new equilibrium would be Y = $5,600.

FIGURE 26-5 THE DEPENDENCE OF NET EXPORTS ON GDP

This graph displays the data on exports, imports, and net exports found in Table 26-4. Exports, *X*, are independent of GDP while imports, *IM*, rise as GDP rises (top panel). As a result, net exports, (*X* – *IM*), decline as GDP rises (bottom panel).

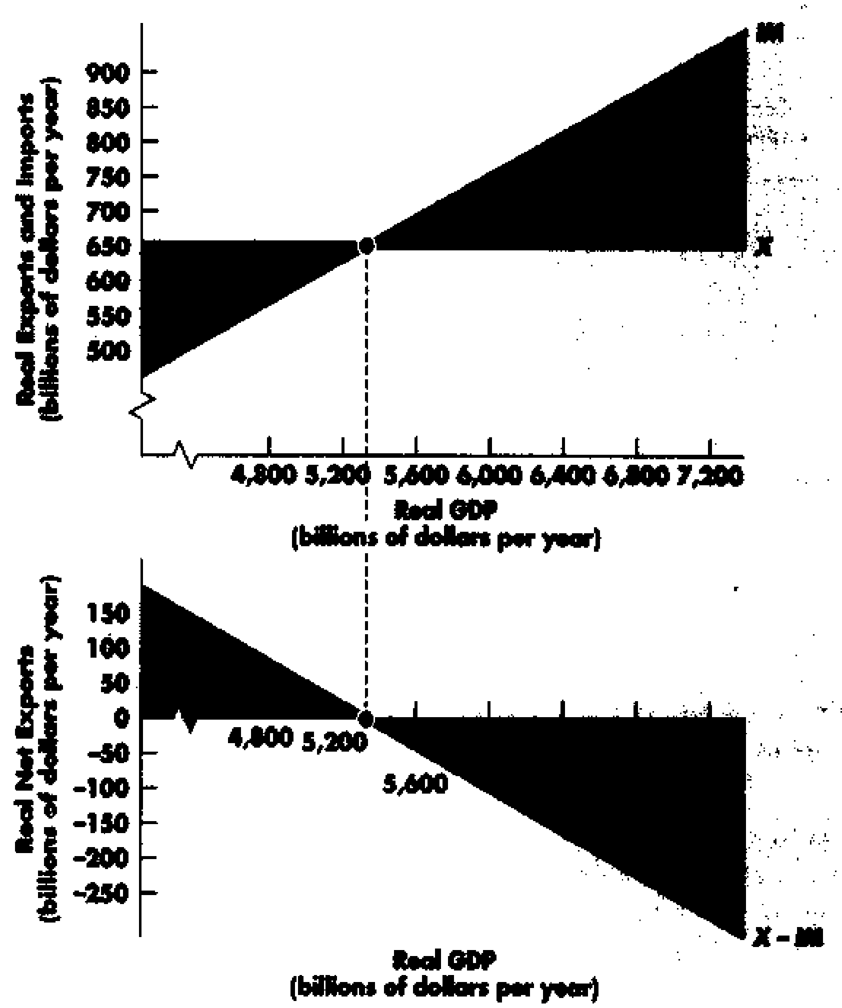


FIGURE 26-6 EQUILIBRIUM GDP WITH VARIABLE IMPORTS

In the presence of variable imports, equilibrium GDP occurs where the orange *C* + *I* + *G* + (*X* – *IM*) line, rather than the black one, crosses the 45° line. In the graph, equilibrium is at point *E*, where GDP is $6,000 billion. This matches the equilibrium we found in Chapter 25 (Figure 25-3 on page 597) with fixed imports because we have designed the example to come out that way.

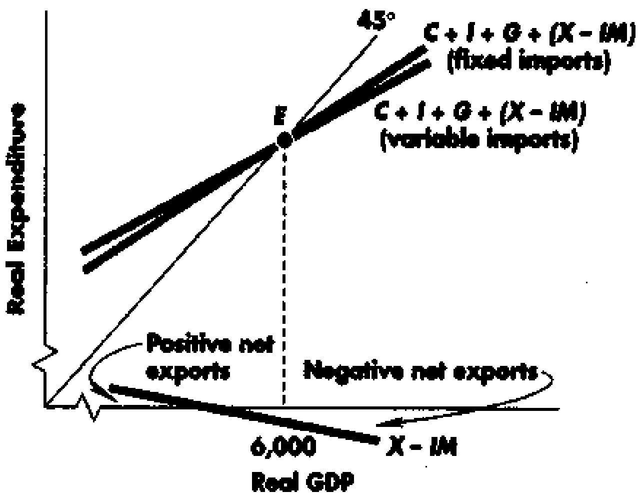


TABLE 26-5 EQUILIBRIUM INCOME AFTER A $160 BILLION RISE IN EXPORTS (BILLIONS OF DOLLARS PER YEAR)

(1) Gross Domestic Product (Y)	(2) Consumer Expenditures (C)	(3) Investment (I)	(4) Government Purchases (G)	(5) Exports (X)	(6) Imports (IM)	(7) Net Exports (X − IM)	(8) Total Expenditure [C + I + G + (X − IM)]
4,800	3,000	900	1,300	810	570	+240	5,440
5,200	3,300	900	1,300	810	630	+180	5,680
5,600	3,600	900	1,300	810	690	+120	5,920
6,000	3,900	900	1,300	810	750	+60	6,160
6,400	**4,200**	**900**	**1,300**	**810**	**810**	**0**	**6,400**
6,800	4,500	900	1,300	810	870	−60	6,640
7,200	4,800	900	1,300	810	930	−120	6,880

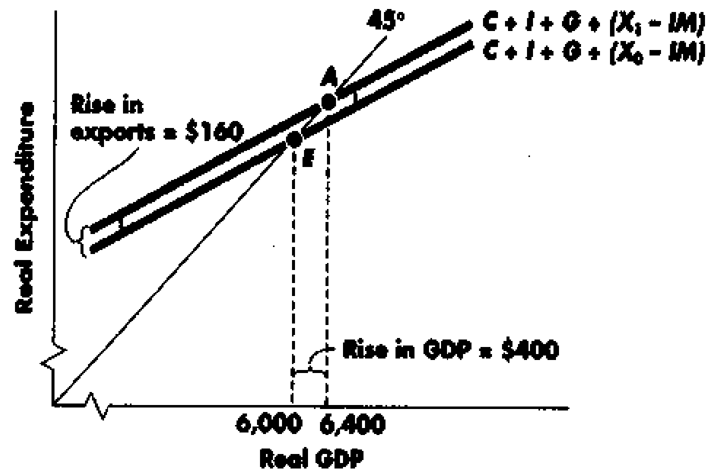
This same conclusion is shown graphically in Figure 26-7, where the line $C + I + G + (X_0 - IM)$ represents the original expenditure schedule and the line $C + I + G + (X_1 - IM)$ represents the expenditure schedule after the rise in exports. Equilibrium shifts from point E to point A, and GDP rises by $400 billion.

Notice that the multiplier in this example is 2.5, whereas in the chapter, with net exports taken to be a fixed number, it was 4. This simple example illustrates a general result: *International trade lowers the numerical value of the multiplier.* Why is this so? Because, in an open economy, any autonomous increase in spending is partly dissipated in purchases of foreign goods, which creates additional income for *foreigners* rather than for domestic citizens.

Figure 26-6 shows this same conclusion graphically. Because net exports decline as GDP rises, the total expenditure line is *flatter* in the presence of variable imports [orange $C + I + G + (X - IM)$ line] than it would be with fixed imports (black

FIGURE 26-7 THE MULTIPLIER WITH VARIABLE IMPORTS

This diagram shows a $160 billion increase in exports as a vertical shift of the total expenditure schedule from $C + I + G + (X_0 - IM)$, to $C + I + G + (X_1 - IM)$. As a result, equilibrium shifts from point E to point A, and GDP rises from $6,000 billion to $6,400 billion. The multiplier is therefore 2.5 (= $400/$160).



line). As we know from earlier chapters, the *size* of the multiplier depends on the *slope* of the expenditure schedule—steeper expenditure schedules lead to larger multipliers. Since variable imports flatten the expenditure schedule, they lower the multiplier.[6]

Thus, international trade gives us the first of what will eventually be several reasons why the oversimplified multiplier formula overstates the true value of the multiplier.

[6] For those who like formulas, we can amend the oversimplified multiplier formula to allow for international trade. That formula was: Multiplier = $1/(1 - b)$ where b is the marginal propensity to consume. If we define the *marginal propensity to import* as the rise in imports per dollar of GDP (the marginal propensity to import is 0.15 in our example) and symbolize it by the letter m, the formula for the multiplier with foreign trade is: Multiplier $= 1/(1 - b + m)$. This formula clearly shows that a higher value of m leads to a lower multiplier.

SUMMARY

1. Because imports rise as GDP rises while exports are insensitive to (domestic) GDP, net exports decline as GDP rises.
2. If imports depend on GDP, international trade reduces the value of the multiplier.

QUESTIONS FOR REVIEW

1. Suppose exports and imports of a country are given by:

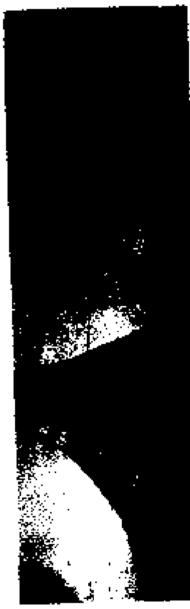
GDP	Exports	Imports
$2,500	$400	$250
3,000	400	300
3,500	400	350
4,000	400	400
4,500	400	450
5,000	400	500

Calculate net exports at each level of GDP.

2. If domestic expenditure (the sum of $C + I + G$ in the economy described in Review Question 1) is as shown below, construct a 45° line diagram and locate the equilibrium level of GDP.

GDP	Domestic Expenditures
$2,500	$3,100
3,000	3,400
3,500	3,700
4,000	4,000
4,500	4,300
5,000	4,600

3. Now raise exports to $650 and find the equilibrium again. How large is the multiplier?



CHAPTER 27

SUPPLY-SIDE EQUILIBRIUM: UNEMPLOYMENT *AND* INFLATION?

We might as well reasonably dispute whether it is the upper or the under blade of a pair of scissors that cuts a piece of paper, as whether value is governed by [demand] or [supply].
Alfred Marshall

In Chapter 25, we learned that the position of the economy's total expenditure schedule governs whether the economy will experience a recessionary or an inflationary gap. If the $C + I + G + (X - IM)$ schedule is "too low," a *recessionary gap* will arise. A $C + I + G + (X - IM)$ schedule that is "too high" leads to an *inflationary gap*. Which sort of gap actually occurs is of considerable importance because a recessionary gap normally spells unemployment while an inflationary gap leads to inflation.

The tools provided in Chapter 25, however, are not sufficient to determine which sort of gap will arise because, as we learned there, the position of the expenditure schedule depends on the price level. And the price level is determined by *both* aggregate demand *and* aggregate supply. Hence, the task of this chapter: to bring the supply side of the economy into the picture.

We begin by explaining how the *aggregate supply curve* is derived from business costs. Next we consider the interaction of aggregate supply and aggregate demand, and how they jointly determine output and the price level. With this apparatus in hand, we return to recessionary and inflationary gaps and study how the economy adjusts to each. Doing this puts us in a position to deal with the crucial question raised in earlier chapters: Does the economy have an efficient self-correcting mechanism? We shall see that the answer is "yes, but." Yes, but it works slowly. Finally, we use aggregate supply–aggregate demand analysis to explain the vexing problem of *stagflation*—the simultaneous occurrence of high unemployment *and* high inflation—that plagued the economy in the 1980s.

DEBATE: TWO SIDES TO THE SUPPLY SIDE

In 1981, President Ronald Reagan brought to Washington a doctrine called *supply-side economics*—a new theory advertised as a replacement for the Keynesian theory we have been studying so far. Reaganite supply-side economics, which was controversial from the start, emphasized tax cuts that allegedly would increase saving and investment—and thus augment the *supply* of capital.

Twelve years later, the American voters repudiated Reaganomics and voted in President Bill Clinton who, ironically, also ran on what might be called a "supply-side" platform. Clintonomics, with its emphasis on upgrading the skills of the American work force through education and training, was starkly different from Reaganomics. But the two programs share one idea in common: that what happens on the *supply* side of the economy greatly affects inflation, unemployment, and economic growth.

This debate rages on as President Clinton and the Republican-controlled Congress battle over budget and tax priorities. The president insists on maintaining key government "investment" programs like job training and education. The Republicans insist on deeper spending cuts to make room for tax cuts. The political impasse over budget priorities actually led to two partial shutdowns of the U.S. government in late 1995.

It is therefore time for us to consider the origins of the aggregate supply curve and the factors that can make it shift. Only when we have done so will we be ready to analyze the *joint* determination of output *and* the price level—with its consequences for inflationary and recessionary gaps—and to render a judgment on competing supply-side agendas. Read the next two chapters before making up your mind.

THE AGGREGATE SUPPLY CURVE

In earlier chapters we noted that aggregate demand is a schedule, not a fixed number. The quantity of real GDP that will be demanded depends on the price level, as summarized in the economy's *aggregate demand curve.*

Analogously, the concept of *aggregate supply* does not refer to a fixed number, but rather to a schedule (a *supply curve*). The volume of goods and services that profit-seeking enterprises will provide depends on the prices they obtain for their outputs, on wages and other production costs, on the state of technology, and on other things. The relationship between the price level and the quantity of real GDP supplied, *holding all other determinants of quantity supplied constant,* is called the economy's **aggregate supply curve.**

The **AGGREGATE SUPPLY CURVE** shows, for each possible price level, the quantity of goods and services that all the nation's businesses are willing to produce during a specified period of time, holding all other determinants of aggregate quantity supplied constant.

A typical aggregate supply curve is drawn in Figure 27-1. It slopes upward, meaning that as prices rise more output is produced, *other things held constant.* It is not difficult to understand why. Producers in the U.S. economy are motivated mainly by profit. The profit made by producing a unit of output is simply the difference between the price at which it is sold and the unit cost of production:

$$\text{Profit per unit} = \text{Price} - \text{Cost per unit}$$

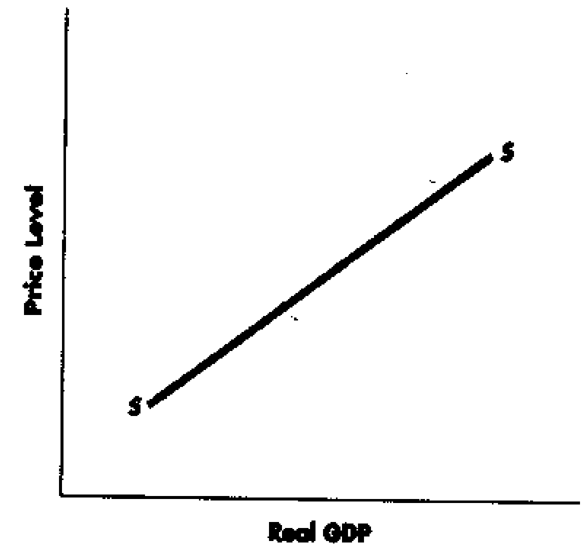
So, the response of output to a rising price level—which is what the slope of the aggregate supply curve shows—depends on the response of costs.

Many of the prices that firms pay for labor and other inputs are relatively fixed for periods of time—though certainly not forever. Often, workers and firms enter into long-term labor contracts that set money wages up to 3 years in advance. Even where there are no explicit contracts, wage rates typically adjust only once a year. During the interim period, money wages are fixed. Similarly, a variety of material inputs are delivered to firms under long-term contracts at prearranged prices.

Why is it significant that firms often purchase inputs at prices that stay fixed for considerable periods of time? Because firms decide how much to produce by

FIGURE 27-1 AN AGGREGATE SUPPLY CURVE

This graph shows a typical aggregate supply curve. It has a positive slope (that is, it rises as we move to the right), meaning that the quantity of output supplied rises as the price level rises.



comparing their selling prices with their costs of production; and production costs depend, among other things, on input prices. If the selling prices of the firm's products rise while its wages and other factor costs are fixed, production becomes more profitable, and firms will presumably increase output.

A simple example will illustrate the idea. Suppose a firm uses 1 hour of labor to manufacture a gadget that sells for $9. If workers earn $8 per hour, and the firm has no other production costs, its profit per unit is:

$$\begin{aligned}\text{Profit per unit} &= \text{Price} - \text{Cost per unit}\\ &= \$9 - \$8 = \$1\end{aligned}$$

Now what happens if the price of a gadget rises to $10, but wage rates remain constant? The firm's profit per unit becomes:

$$\begin{aligned}\text{Profit per unit} &= \text{Price} - \text{Cost per unit}\\ &= \$10 - \$8 = \$2\end{aligned}$$

With production more profitable, the firm will likely supply more gadgets.

The same process operates in reverse. If selling prices fall while input costs are relatively fixed, profit margins will be squeezed and production cut back. This behavior is summarized by the upward slope of the aggregate supply curve: Production rises when the price level (henceforth, P) rises, and falls when P falls. In other words:

The aggregate supply curve slopes upward because firms normally can purchase labor and other inputs at prices which are fixed for some period of time. Thus, higher selling prices for output make production more attractive.[1]

---

[1]There are both differences and similarities between the *aggregate* supply curve and the *microeconomic* supply curves studied in Chapter 4. Both are based on the idea that quantity supplied depends on how output prices move relative to input prices. But the aggregate supply curve pertains to the behavior of *the overall price level*, whereas a microeconomic supply curve pertains to the *price of some particular commodity*.

The phrase "for some period of time" alerts us to an important fact: The aggregate supply curve may not stand still for long. If wages or prices of other inputs change, as they surely will during inflationary times, then the aggregate supply curve will shift.

SHIFTS OF THE AGGREGATE SUPPLY CURVE

We have concluded so far that, for given levels of wages and other input prices, there will be an upward-sloping aggregate supply curve relating the price level to aggregate quantity supplied. Now let us consider what happens when these input prices change.

THE MONEY WAGE RATE

The most obvious determinant of the position of the aggregate supply curve is the *money wage rate*. Wages are the major element of cost in the economy, accounting for more than 70 percent of all inputs. Since higher wage rates mean higher costs, they spell lower profits at any given prices. That is why companies like American Airlines and Caterpillar have staged fierce battles with their unions in recent years in an effort to reduce wages.

Returning to our example, consider what would happen to a gadget producer if the money wage rose to $8.75 per hour while the price of a gadget remained $9. Profit per unit would decline from:

$$\$9 - \$8 = \$1$$

to

$$\$9.00 - \$8.75 = \$0.25$$

With profits squeezed, the firm would probably cut back on production.

This is the way firms in our economy typically react to a rise in wages. Therefore, a wage increase leads to a decrease in aggregate quantity supplied at current prices. Graphically, the aggregate supply curve shifts to the left (or inward), as shown in Figure 27-2. In this diagram, firms are willing to supply $6,000 billion in goods and services at a price level of 100 when wages are low (point *A*). But after wages increase these same firms are willing to supply only $5,500 billion at this price level (point *B*). By similar reasoning, the aggregate supply curve will shift to the right (or outward) if wages fall. Thus:

A rise in the money wage rate makes the aggregate supply curve *shift inward*, meaning that the quantity supplied at any price level *declines*. A fall in the money wage rate makes the aggregate supply curve *shift outward*, meaning that the quantity supplied at any price level *increases*.

PRICES OF OTHER INPUTS

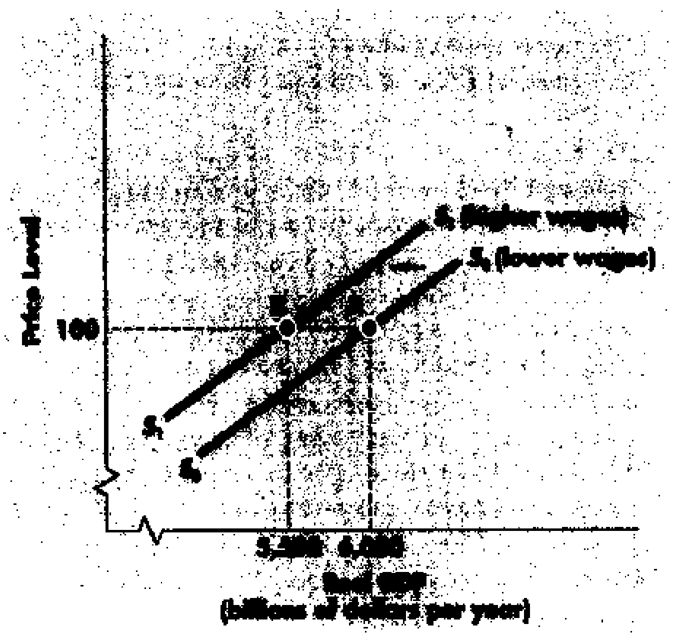
In this regard, there is nothing special about wages. An increase in the price of *any* input that firms buy will shift the aggregate supply curve in the same way; that is:

The aggregate supply curve is shifted inward by an increase in the price of any input to the production process, and it is shifted *outward* by any decrease.

While there are many inputs other than labor, the one that has attracted the most attention in recent decades is energy. Increases in the price of energy, such

FIGURE 27-2

A SHIFT OF THE AGGREGATE SUPPLY CURVE



This diagram shows what happens to the economy's aggregate supply curve when money wages rise. Higher wages shift the supply curve inward from S_0S_0 to S_1S_1, leading, for example, to an output level of \$5,500 billion (point *B*), rather than \$6,000 (point *A*), when the price level is 100. The aggregate supply curve will shift inward in the same manner if the price of any other input (such as energy) increases.

as those that took place in the early 1980s and again during the 1990 Gulf war, push the aggregate supply curve inward more or less as shown in Figure 27-2. By the same token, a rise in the price of *any* input we import from abroad would have the effect shown in the figure.

TECHNOLOGY AND PRODUCTIVITY

PRODUCTIVITY is the amount of output produced by a unit of input.

Another factor that determines the position of the aggregate supply curve is the state of technology. Suppose, for example, that a technological breakthrough increases the **productivity** of labor, that is, output per hour of work. If wages do not change, such an improvement in productivity will *decrease* business costs, improve profitability, and encourage more production.

Once again, our gadget company will help us understand how this works. Suppose the price of a gadget stays at \$9 and the hourly wage rate stays at \$8, but gadget workers become much more productive. Specifically, suppose the labor input required to manufacture a gadget falls from 1 hour (which costs \$8) to three-quarters of an hour (which costs \$6). Then profit per unit rises from

$$\$9 - \$8 = \$1$$

to

$$\$9 - \$6 = \$3$$

The lure of higher profits should induce gadget manufacturers to increase production—which is, of course, why manufacturers are constantly striving to raise productivity. In brief, we have concluded that:

Improvements in productivity shift the aggregate supply curve outward.

Figure 27-2 can therefore be viewed as applying to a *decline* in productivity. Since the 1970s, slow growth of productivity has been a persistent problem for the U.S. economy, one that we will examine in depth in Chapter 37.

AVAILABLE SUPPLIES OF LABOR AND CAPITAL

The last determinant of the position of the aggregate supply curve is obvious. The bigger the economy—as measured by its available supplies of labor and capital—the more it is capable of producing. So:

As the labor force grows or improves in quality, and as the capital stock is increased by investment, the aggregate supply curve shifts *outward* to the right, meaning that more output can be produced at any given price level.

This last aspect of the aggregate supply curve is central to the political debate over alternative supply-side strategies. Although neither party excludes the other factor of production, Republicans tend to concentrate on augmenting the supply of *capital* while Democrats tend to emphasize improvements in *labor* quality.

These, then, are the major "other things" that we hold constant when drawing up an aggregate supply curve: wage rates, prices of other inputs (such as energy), technology, labor force, and capital stock. While a change in the price level moves the economy *along a given supply curve*, a change in any of the other determinants of aggregate quantity supplied *shifts the entire supply schedule.*

EQUILIBRIUM OF AGGREGATE DEMAND AND SUPPLY

Chapter 25 taught us that the price level is a crucial determinant of whether equilibrium GDP is below full employment (a "recessionary gap"), precisely at full employment, or above full employment (an "inflationary gap"). We are now in a position to analyze which type of gap, if any, will actually occur in any particular case. By combining the analysis of aggregate supply just completed with the analysis of aggregate demand from the last two chapters, we can determine *simultaneously* the equilibrium level of real GDP (Y) and the equilibrium price level (P).

Figure 27-3 displays the mechanics. Aggregate demand curve DD and aggregate supply curve SS intersect at point E, where real GDP is \$6,000 billion and the price level is 100. As can be seen in the graph, at any higher price level, such as 120, aggregate quantity supplied would exceed aggregate quantity demanded. There would be a glut on the market as firms found themselves unable to sell all their output. As inventories piled up, firms would compete more vigorously for the available customers, thereby forcing prices down. Both the price level and production would fall.

At any price level lower than 100, such as 80, quantity demanded would exceed quantity supplied. There would be a shortage of goods on the market. With inventories disappearing and customers knocking on their doors, firms would be encouraged to raise prices. The price level would rise, and so would output. Only when the price level is 100 are the quantities of real GDP demanded and supplied equal. Therefore, only the combination of $P = 100$, $Y = \$6,000$ is an equilibrium.

Table 27-1 illustrates the same conclusion in another way, using a tabular analysis similar to that of Chapter 25 (refer back to Table 25-2, page 597). Columns 1 and 2 constitute an aggregate demand schedule corresponding to the aggregate demand curve DD in Figure 27-3. Columns 1 and 3 constitute an aggregate supply schedule corresponding exactly to aggregate supply curve SS in the figure.

It is clear from the table that equilibrium occurs only at $P = 100$ and $Y = \$6,000$. At any other price level, aggregate quantities supplied and demanded

FIGURE 27-3 EQUILIBRIUM OF REAL GDP AND THE PRICE LEVEL

This diagram shows how the equilibrium levels of real GDP and the price level are simultaneously determined by the intersection of the aggregate demand curve (*DD*) and the aggregate supply curve (*SS*). In this example, equilibrium occurs at point *E*, with a real GDP of $6,000 billion and a price level of 100.

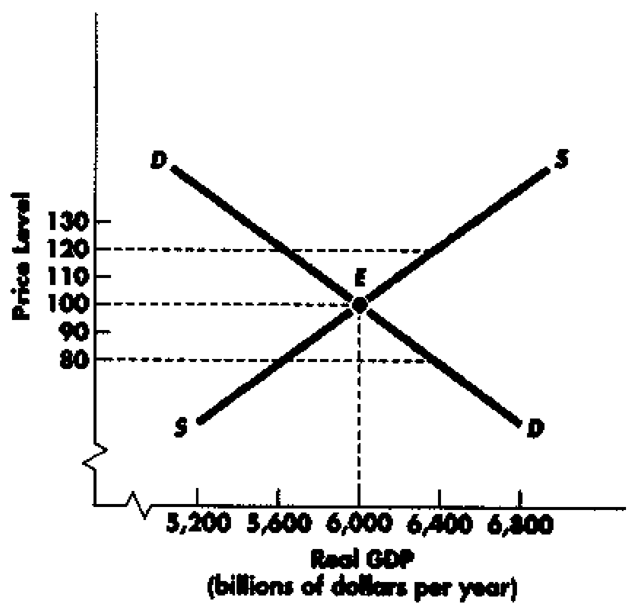


TABLE 27-1 DETERMINATION OF THE EQUILIBRIUM PRICE LEVEL

(1) Price Level	(2) Aggregate Quantity Demanded (billions of dollars)	(3) Aggregate Quantity Supplied (billions of dollars)	(4) Balance of Supply and Demand	(5) Prices Will
80	6,400	5,600	Quantity demanded exceeds quantity supplied	Rise
90	6,200	5,800	Quantity demanded exceeds quantity supplied	Rise
100	**6,000**	**6,000**	**Quantity demanded equals quantity supplied**	**Remain the same**
110	5,800	6,200	Quantity supplied exceeds quantity demanded	Fall
120	5,600	6,400	Quantity supplied exceeds quantity demanded	Fall

would be unequal, with consequent upward or downward pressure on prices. For example, at a price level of 90, customers demand $6,200 billion worth of goods and services, but firms wish to provide only $5,800 billion. The price level is too low and will be forced upward. Conversely, at a price level of, say, 110, quantity supplied ($6,200 billion) exceeds quantity demanded ($5,800 billion), implying that the price level must fall.

RECESSIONARY AND INFLATIONARY GAPS REVISITED

Let us now reconsider a question we posed, but could not answer, in Chapter 25: Will equilibrium occur at, below, or beyond full employment?

We could not give a complete answer to this question in Chapter 25 because we had no way to determine the equilibrium price level, and therefore no way to tell which type of gap, if any, would arise. The aggregate supply and demand analysis summarized in Figure 27-3 now gives us what we need. But we find that our answer is the same as it was in Chapter 25—anything can happen.

The reason is that nothing in Figure 27-3 tells us where full employment is; it could be above the $6,000 billion equilibrium level or below it. Depending on the locations of the aggregate demand and aggregate supply curves, then, we can reach equilibrium *above* full employment (an inflationary gap), *at* full employment, or *below* full employment (a recessionary gap). In the short run, with wages and other input costs fixed, that is all there is to it.

All three possibilities are illustrated in Figure 27-4. The three upper panels are familiar from Chapter 25. As we move from left to right, the expenditure schedule rises from $C + I_0 + G + (X - IM)$ to $C + I_1 + G + (X - IM)$ to $C + I_2 + G + (X - IM)$, leading respectively to a recessionary gap, an equilibrium at full employment, and an inflationary gap. In fact, the upper left-hand diagram looks just like Figure 25-6 (page 601), and the upper right-hand diagram duplicates Figure 25-7 (page 602). We emphasized in Chapter 25 that any one of the three cases is possible, depending on the price level and the expenditure schedule.

In the three lower panels, the equilibrium price level is determined at point E by the intersection of the aggregate supply curve (SS) and the aggregate demand curve (DD). But the same three possibilities emerge.

In the lower left-hand panel, aggregate demand is too low to provide jobs for the entire labor force, so there is a recessionary gap equal to distance EB, or $1,000 billion. This corresponds precisely to the situation depicted on the income-expenditure diagram immediately above it.

In the lower right-hand panel, aggregate demand is so high that the economy reaches an equilibrium well beyond full employment. There is an inflationary gap equal to BE, or $1,000 billion, just as in the diagram immediately above it.

In the lower middle panel, the aggregate demand curve D_1D_1 is at just the right level to produce an equilibrium at full employment. There is neither an inflationary nor a recessionary gap, as in the diagram just above it.

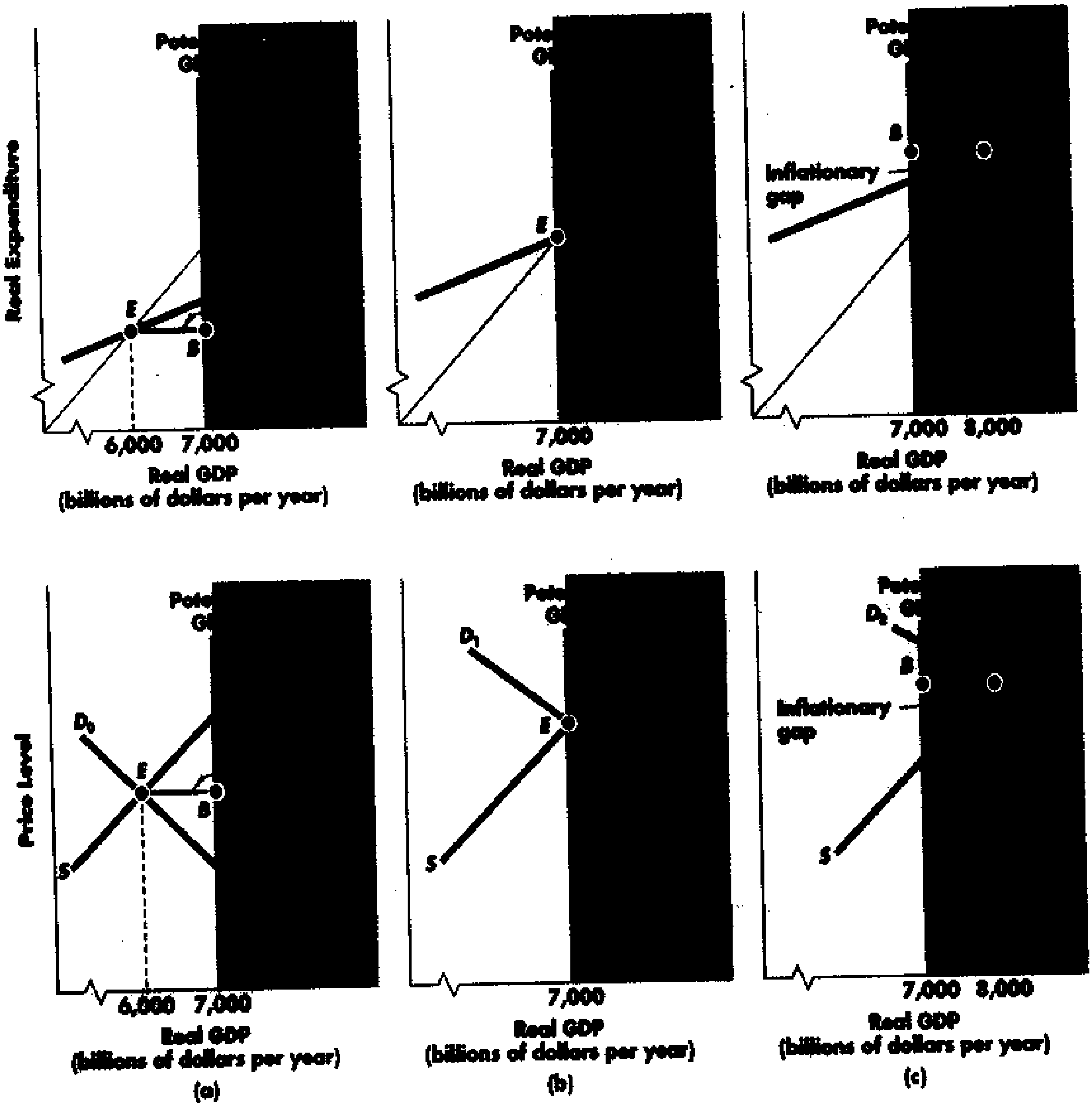
It may seem, therefore, that we have done nothing but restate our previous conclusions. But, in fact, we have done much more. Because now that we have studied the determination of the equilibrium price level, we are able to examine how the economy adjusts to either a recessionary gap or an inflationary gap. Specifically, since wages are fixed in the short run, any one of the three cases depicted in Figure 27-4 can occur. But, in the long run, wages will adjust to labor market conditions. It is to that adjustment that we now turn.

ADJUSTING TO A RECESSIONARY GAP: DEFLATION OR UNEMPLOYMENT?

Suppose the economy starts with a recessionary gap—that is, an equilibrium *below* full employment—as in the lower left-hand panel of Figure 27-4. This might be caused, for example, by inadequate consumer spending or by anemic investment spending. What happens next?

With equilibrium GDP below potential, jobs will be hard to find. The ranks of the unemployed will exceed the number expected to be jobless because of moving, changing occupations, and so on. In the terminology of Chapter 23,

FIGURE 27-4 RECESSIONARY AND INFLATIONARY GAPS REVISITED



This diagram shows three possible types of equilibrium on two different diagrams. In the top row, income-expenditure diagrams from Chapter 25 are used to depict a recessionary gap, an equilibrium at full employment, and an inflationary gap. In the bottom row, these same three situations are shown on aggregate supply and demand diagrams. In each case, the aggregate supply curve is the same (SS), equilibrium occurs at point E, and full-employment GDP is $7,000 billion. In Panel (a), the aggregate demand curve D_0D_0 is relatively low, so that equilibrium falls below full employment. There is a recessionary gap measured by the distance EB, or $1,000 billion. In Panel (b), the aggregate demand curve D_1D_1 is higher, and equilibrium occurs precisely at full employment. There is no gap of either kind. In Panel (c), the aggregate demand curve D_2D_2 is so high that equilibrium occurs beyond full employment. There is an inflationary gap measured by the distance BE, or $1,000 billion.

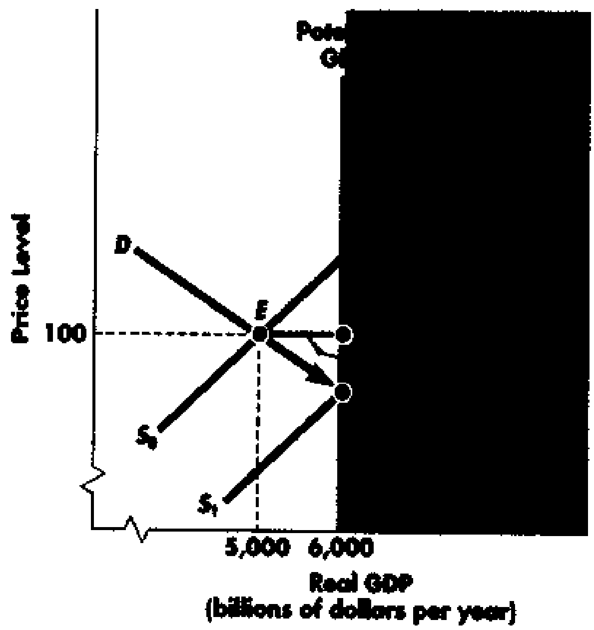
there will be a considerable amount of *cyclical unemployment*. Businesses, on the other hand, will have little trouble finding workers. And their current employees will be eager to hang on to their jobs.

In such an environment, it will be very difficult for workers to win wage increases. Indeed, in extreme situations, wages may even fall—thus shifting the aggregate supply curve *outward*. (Remember, an aggregate supply curve is drawn for a *given* money wage.) But as the aggregate supply curve shifts outward—eventually moving from S_0S_0 to S_1S_1 in Figure 27-5—prices decline and the recessionary gap shrinks. This is the process by which deflation erodes the recessionary gap, eventually leading the economy to an equilibrium at full employment (point *F* in Figure 27-5).

But there is an important catch. In our modern economy, this adjustment process proceeds slowly—painfully slowly. Our brief review of the historical record in Chapter 22 showed that the history of the United States includes several examples of *deflation* before World War II but none since. Not even the severe recession of 1981–1982, during which unemployment climbed above 10 percent, was able to force average prices and wages down—though it certainly slowed their rates of increase. Similarly, the recession of the early 1990s reduced inflation, but certainly did not bring deflation.

Exactly why wages and prices rarely fall in our modern economy has been a subject of intense and continuing controversy among economists for years. Some economists emphasize institutional factors like minimum wage laws, union contracts, and a variety of government regulations that place legal floors under particular wages and prices. Because most of these institutions are of relatively recent vintage, this theory successfully explains why wages and prices fall less frequently now than they did before World War II. However, only a small minority of the U.S. economy is subject to legal restraints on wage and price cutting. So it seems doubtful that legal restrictions can provide a complete explanation.

FIGURE 27-5 THE ELIMINATION OF A RECESSIONARY GAP



At point *E* there is a recessionary gap because the aggregate demand curve *DD* crosses the aggregate supply curve S_0S_0 below the level of potential GDP. As wages fall, the aggregate supply curve gradually shifts outward until it reaches the position indicated by supply curve S_1S_1. Here the economy has attained a full-employment equilibrium at point *F*. But if wages fall very slowly, the economy gets stuck with a recessionary gap and high unemployment for a long time.

Other observers suggest that workers have a profound psychological resistance to accepting a wage reduction. This theory certainly has the ring of truth. Think how you would react if your boss announced he was cutting your hourly wage rate. You might quit, or you might devote less care to your job. If the boss suspects you will react this way, he may be reluctant to cut your wage. Nowadays, genuine wage reductions are rare enough to be newsworthy. United Airlines was one of the few firms that managed to cut wages; workers accepted an average 15 percent pay cut shortly after a 1994 employee buyout of the company. Many other companies have tried, but failed, to win similar concessions from their workers.

While no one doubts that wage cuts can be bad for morale, the psychological theory has one major drawback. It fails to explain why the psychological resistance to wage cuts apparently started only after World War II. Until a satisfactory answer to this question is provided, many economists will remain skeptical.

A third explanation is based on a fact we emphasized in Chapter 22—that business cycles were less severe in the postwar period than they were in the prewar period. Because workers and firms came to believe that recessions would not turn into depressions, the argument goes, they may have decided to wait out the bad times rather than accept wage or price reductions that they would later regret.

Yet another theory is based on the old adage, "you get what you pay for." The idea is that workers differ in productivity, but that productivities of individual employees are hard to identify. Firms therefore worry that a general wage reduction will result in the loss of their best employees—since these workers have the best opportunities elsewhere in the economy. Rather than take this chance, the argument goes, firms prefer to maintain high wages even in recessions.

There are other theories as well, none of which commands a clear majority of professional opinion. But, regardless of the cause, we may as well accept the fact that, in our modern economy, prices and wages generally fall only sluggishly when demand is weak.

The implications of this rigidity are quite serious, for a recessionary gap cannot cure itself without some deflation. And if wages and prices will not fall, recessionary gaps like *EB* in Figure 27-5 will linger for a long time. That is:

When aggregate demand is low, the economy may get stuck with a recessionary gap for a long time. If wages and prices fall very slowly, the economy will endure a prolonged period of production below potential GDP.

DOES THE ECONOMY HAVE A SELF-CORRECTING MECHANISM?

Now a situation like this would, presumably, not last forever. As the recession lengthened, and perhaps deepened, more and more workers would be unable to find jobs at the prevailing high wages. Eventually, their need to be employed would overwhelm their resistance to wage cuts.

Firms, too, would become increasingly willing to cut prices as the period of weak demand persisted and managers became convinced that the slump was not merely a temporary aberration. Prices and wages did, in fact, fall during the Great Depression of the 1930s. And some fell again in the weak markets of the early 1990s.

However, nowadays political leaders of both parties believe it is folly to wait for falling wages and prices to eliminate a recessionary gap. They agree that *some* government action is both necessary and appropriate under recessionary

TIGHT LABOR MARKETS IN 1995

When the unemployment rate fell to around 5.5 percent in 1995, jobs became hard to fill in certain parts of the country. The National Federation of Independent Business reported that personnel managers encountered difficulty in filling one out of five corporate job openings. A summary business report prepared by the 12 Federal Reserve banks described the labor market as *tight* and noted wage increases in selected occupations.

Even wages for unskilled workers in retail businesses got a boost. Since 1991, hourly wages for retail help had increased only 3 percent annually. But the low unemployment rate changed things. McDonald's, which normally pays $5.00 per hour, started advertising jobs starting at $6.50 per hour in places like Winston-Salem, North Carolina—a 30 percent increase. The situation was similar for cashiers and clerks in Detroit, where employers had to raise pay rates 10 percent to fill positions.

To firms, it all meant higher labor costs, and, as a result, the aggregate supply curve shifted inward. But for workers accustomed to living under the specter of unemployment, it was the best news in years.

SOURCES: David Hage, "Have Job Prospects Started to Wilt?" *U.S. News and World Report*, May 15, 1995, p. 64; Vivian Brownstein and Joseph Spiers, "The Growing Threat of Inflation," *Fortune*, January 16, 1995, p. 66; and "New Threats to the Expansion," *Fortune*, July 25, 1994, p. 56.

the size of the inflationary gap declines. In other words, inflation eventually erodes the inflationary gap and brings the economy to an equilibrium at full employment (point *F*).

There is a straightforward way of looking at the economics that underlies this process. Inflation arises because buyers are demanding more output than the economy is capable of producing at normal operating rates. To paraphrase an old cliché, there is too much demand chasing too little supply. Such an environment encourages price hikes.

But rising prices eat away at the purchasing power of consumers' wealth, forcing them to cut back on consumption, as explained in Chapter 24. In addition, exports fall and imports rise, as we learned in Chapter 25. Eventually, aggregate quantity demanded is scaled back to the economy's capacity to produce; and, at this point, the self-correcting process stops. That, in essence, is the unhappy process by which the economy cures itself of the excessive aggregate demand. In brief:

If aggregate demand is exceptionally high, the economy may reach a short-run equilibrium above full employment (an *inflationary gap*). When this occurs, the tight situation in the labor market soon forces wages to rise. Since rising wages increase business costs, prices increase; there is inflation. As higher prices cut into consumer purchasing power and net exports, the inflationary gap begins to close.

As the inflationary gap is closing, output falls and prices continue to rise; so the economy experiences *stagflation* until the gap is eliminated. At this point, a long-run equilibrium is established with a higher price level and with GDP equal to potential GDP.

Two caveats about this process should be mentioned. One we have already emphasized: The self-correcting mechanism takes time because wages and prices do not adjust quickly. An inflationary gap sows the seeds of its own destruction, but these seeds germinate slowly. So, once again, policymakers may want to speed up the process.

The other caveat is that the process works *only in the absence of additional forces propelling the aggregate demand curve outward*. But in the last chapter, we encountered several forces that might shift the aggregate demand curve outward.

As you can see by manipulating the aggregate demand–aggregate supply diagram, if aggregate demand is shifting out at the same time that aggregate supply is shifting in, there will certainly be inflation, but the inflationary gap may not shrink. (Try this as an exercise, to make sure you understand how to use the apparatus.) So not all inflations come to a natural end.

DEMAND INFLATION AND STAGFLATION

Simple as it is, this adjustment model teaches us a number of important lessons about inflation in the real world. First, Figure 27-6 reminds us that the real culprit in this particular inflation is excessive aggregate demand—relative to potential GDP. The aggregate demand curve is initially so high that it intersects the aggregate supply curve well beyond full employment. The resulting intense demand for goods and labor pushes prices and wages higher. While aggregate demand in excess of potential GDP is not the only possible cause of inflation, it certainly is the cause in our example.

However, business managers and journalists may blame inflation on rising wages. In a superficial sense, of course, they are right, because higher wages do indeed lead firms to raise their prices. But in a deeper sense they are wrong. Both rising wages and rising prices are symptoms of an underlying malady: too much aggregate demand. Blaming labor for inflation in such a case is a bit like blaming high doctor bills for making you ill.

Second, we see that output falls while prices rise as the economy adjusts from point *E* to point *F* in Figure 27-6. This is our first (but not our last!) explanation of the phenomenon of **stagflation.** Specifically:

STAGFLATION is inflation that occurs while the economy is growing slowly or having a recession.

A period of stagflation is part of the normal aftermath of a period of excessive aggregate demand.

It is easy to understand why stagflation occurs in this case. When aggregate demand is excessive, the economy will temporarily produce beyond its normal capacity. Labor markets tighten and wages rise. Machinery and raw materials may also become scarce and so start rising in price. Faced by higher costs, the natural reaction of business firms is to produce less and to charge a higher price. That is stagflation.

TWO RECENT EXAMPLES

The stagflation that follows a period of excessive aggregate demand is, you will note, a rather benign form of the dreaded disease. After all, while output is falling, it nonetheless remains above potential GDP, and unemployment is low. The U.S. economy has experienced two such episodes in the last decade.

The more notable one came between 1988 and 1990. The long economic expansion of the 1980s brought the unemployment rate down to 5.5 percent by mid-1988 and (briefly) to a 15-year low of 5.0 percent by March 1989. Almost all economists believe that 5.0 percent is below the full-employment unemployment rate, that is, that the U.S. economy had an *inflationary gap* in 1989. As the theory suggests, inflation began to accelerate—from 4.4 percent in 1988 to 4.6 percent in 1989 and 6.1 percent in 1990.

In the meantime, the economy was stagnating. Real GDP growth fell from 3.5 percent during 1988 to 2.4 percent in 1989 and −0.2 percent in 1990. Inflation was eating away at the inflationary gap, which was virtually gone by mid-1990,

when the recession started. Yet inflation remained high through the early months of the recession. The U.S. economy was in the *stagflation* phase.

A milder version of this same phenomenon occurred in the first half of 1995. After the unemployment rate fell to 5.4 percent in late 1994, slightly below the full-employment level, the U.S. economy slowed abruptly in the first half of 1995. But inflation during the first half of 1995 ran at a 3.2 percent annual rate, slightly above the 1994 rate of 2.7 percent. In both of these episodes, then, the U.S. economy behaved more or less as our simple model suggests.

Our overall conclusion about the economy's ability to right itself seems to run something like this:

The economy does indeed have a self-correcting mechanism that tends to eliminate either unemployment or inflation. However, this mechanism works slowly and unevenly. In addition, its beneficial effects on either inflation or unemployment are sometimes swamped by strong forces (such as rapid increases or decreases in aggregate demand) pushing in the opposite direction. Thus, the self-correcting mechanism cannot always be relied upon.

STAGFLATION FROM SUPPLY SHIFTS

We have just discussed the type of stagflation that follows in the aftermath of an inflationary boom. However, that is not what happened in the 1970s and early 1980s when unemployment and inflation soared at the same time. What caused this more virulent type of stagflation? Several things, but the principal villain was the rising price of energy.

In 1973, the Organization of Petroleum Exporting Countries (OPEC) quadrupled the price of crude oil. American consumers soon found the prices of gasoline and home heating fuels increasing sharply, and American businesses found that one important cost of doing business—energy prices—rose drastically. OPEC struck again in 1979 to 1980, this time doubling the price of oil. Then the same thing happened a third time, albeit on a smaller scale, when Iraq invaded Kuwait in 1990.

Higher energy prices, we observed earlier, shift the economy's aggregate supply curve *inward* in the manner shown in Figure 27-2 (page 631). If the aggregate supply curve shifts inward, as it surely did in 1973 to 1974, 1979 to 1980, and 1990, production will decline. And in order to reduce demand to the available supply, prices will have to rise. The result is the worst of both worlds: falling production and rising prices.

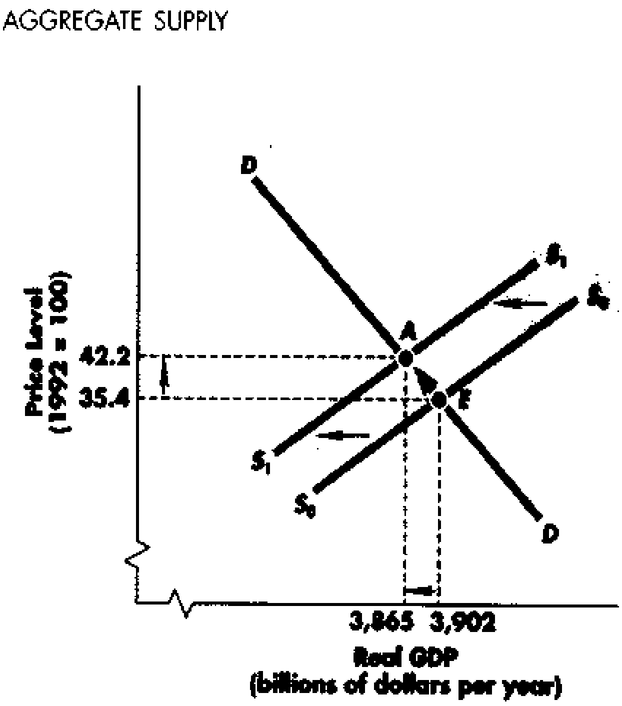
This conclusion is shown graphically in Figure 27-7, which superimposes an aggregate demand curve, *DD*, on the two aggregate supply curves of Figure 27-2. The economy's equilibrium shifts upward to the left, from point *E* to point *A*. Thus, output falls while prices rise. In brief:

Stagflation is the typical result of adverse supply shifts.

The numbers used in Figure 27-7 are roughly indicative of what happened in the United States after the big "energy shock" of late 1973. Between 1973 (represented by supply curve S_0S_0 and point *E*) and 1975 (represented by supply curve S_1S_1 and point *A*), real GDP fell by about 1 percent, while the price level rose a stunning 19 percent. Thus, inflation soared and the economy weakened. The general lesson to be learned from the U.S. experience with supply shocks is both clear and important:

FIGURE 27-7 STAGFLATION FROM AN ADVERSE SHIFT IN AGGREGATE SUPPLY

This diagram illustrates how stagflation arises if the aggregate supply curve shifts inward to the left (from S_0S_0 to S_1S_1). If the aggregate demand curve does not change, equilibrium moves from point *E* to point *A*. Output falls as prices rise, which is what we mean by *stagflation*. The diagram indicates roughly what happened in the United States from 1973 to 1975, when sharply higher energy prices caused severe stagflation.



The typical results of an adverse supply shock are a fall in output and an acceleration in inflation. This is one reason why the world economy was plagued by stagflation in the mid-1970s and early 1980s. And it can happen again if another series of supply-reducing events takes place.

Of course, supply shifts can work in the other direction as well. When the world oil market weakened in 1986, oil prices plummeted. And the price of oil tumbled again just after the Persian Gulf war. Both of these favorable supply shocks stimulated U.S. economic growth and curbed inflation. In fact the Consumer Price Index actually fell for a few months in 1986! The aggregate supply curve was shifting outward.

Favorable supply shocks tend to push output up and reduce inflation.

INFLATION AND THE MULTIPLIER

When we introduced the concept of the multiplier in the last chapter, we said that there were several reasons why its actual value is smaller than suggested by the oversimplified multiplier formula. One of these—variable imports—emerged in an appendix to the last chapter. We are now in a position to understand the second:

Inflation reduces the size of the multiplier.

The basic idea is simple. In the last chapter, we described a multiplier process in which one person's spending becomes another person's income, which leads to further spending by the second person, and so on. But this story is confined to the demand side of the economy. Let us therefore consider what is likely to happen on the supply side as the multiplier process unfolds. Will the additional demand be taken care of by firms without raising prices?

If the aggregate supply curve is upward sloping, the answer is no. More goods will be provided only at higher prices. Thus, as the multiplier chain progresses, pulling income and employment up, prices will also be rising. And this, as we know from earlier chapters, will reduce net exports and dampen consumer spending because rising prices erode the purchasing power of consumers' wealth. So the multiplier chain will not proceed as far as it would have in the absence of inflation.

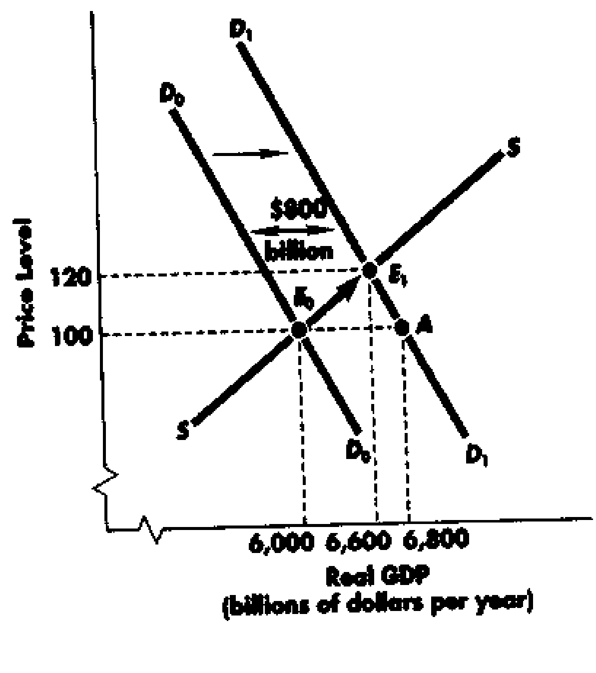
How much inflation results from the rise in demand? How much of the multiplier chain is cut off by inflation? The answers depend on the slope of the economy's aggregate supply curve.

For a concrete example, let us return to the $200 billion increase in investment spending used in the last chapter. There we found (see especially page 620) that $200 billion in additional investment spending eventually leads—through the multiplier process—to *a horizontal shift of $800 billion in the aggregate demand curve.* But to know the actual quantity that will ultimately be produced and the actual price level, we must bring the aggregate supply curve into the picture.

Figure 27-8 does this. Here we show the $800 billion horizontal shift of the aggregate demand curve, from D_0D_0 to D_1D_1, that is derived from the oversimplified multiplier formula (which ignores rising prices). The aggregate supply curve, *SS*, then tells us how this expansion of demand is apportioned between higher output and higher prices. We see that as the economy's equilibrium moves from point E_0 to point E_1, real GDP does not rise by $800 billion. Instead, prices rise, which, as we know, cancels out part of the rise in quantity demanded. So output increases only from $6,000 billion to $6,600 billion—an increase of $600 billion. Thus, in our example, inflation reduces the multiplier from $800/$200 = 4 to $600/$200 = 3. In general:

As long as the aggregate supply curve is upward sloping, any increase in aggregate demand will push up the price level. This, in turn, will drain off some of

FIGURE 27-8 INFLATION AND THE MULTIPLIER



This figure illustrates the complete multiplier analysis, including inflation. The simple multiplier (which ignored changes in the price level) appears as a *horizontal* shift of $800 billion in the aggregate demand curve, meaning that the multiplier would be $800/$200 = 4 if prices did not rise. However, when aggregate demand shifts from D_0D_0 to D_1D_1, the price level increases from 100 to 120. So real GDP increases from $6,000 billion to only $6,600 billion—for a multiplier of $600/$200 = 3.

the higher real demand by eroding the purchasing power of consumer wealth and by reducing net exports. Thus, inflation reduces the value of the multiplier below that suggested by the oversimplified formula.

Notice also that the price level in this example has been pushed up (from 100 to 120, or 20 percent) by the rise in investment demand. This, too, is a general result:

As long as the aggregate supply curve is upward sloping, any outward shift of the aggregate demand curve will cause some rise in prices.

The economic behavior behind these results certainly cannot be considered surprising. Firms faced with large increases in quantity demanded at their original prices respond to these changed circumstances in two natural ways: They raise production (so GDP rises), and they raise prices (so the price level rises). But this rise in the price level reduces the purchasing power of the bank accounts and bonds held by consumers, and they too react in the natural way: They cut down on their spending. Such a reaction amounts to a movement *along* aggregate demand curve D_1D_1 in Figure 27-8 from point A to point E_1.

Higher prices thus play their usual dual role in a market economy: They encourage suppliers to produce more, and they encourage demanders to consume less. In this way, equilibrium is reestablished at higher levels of output and higher prices through the process of inflation.

Figure 27-8 also shows us exactly where the oversimplified multiplier formula goes wrong. By ignoring the effects of the higher price level, the oversimplified formula erroneously supposes that the economy moves horizontally from point E_0 to point A. As the diagram clearly shows, output does not actually rise this much. Output *would* rise this much only if the aggregate supply curve were horizontal. (Verify this for yourself by penciling in an imaginary horizontal aggregate supply curve through points E_0 and A in Figure 27-8.) That is, the oversimplified multiplier formula tacitly assumes that the aggregate supply curve is horizontal. Normally, this is an unrealistic assumption, which is one reason why the oversimplified formula exaggerates the size of the multiplier.

In summary, it may be useful to put together what we have learned about multiplier analysis in this chapter and the last:

STEPS IN CALCULATING THE MULTIPLIER

1. Shift the expenditure schedule in the 45° line diagram vertically by the amount of the autonomous shift in spending (as, for example, in Figure 26-1 on page 611).
2. Use the 45° line diagram, or the oversimplified multiplier formula, to calculate the multiplier effect on GDP that *would* occur *if* the price level did not change (again, see Figure 26-1).
3. Now move from the 45° line diagram to an aggregate supply and demand diagram such as Figure 27-8 to see how the price level will react. Enter the multiplier effect calculated in Step 2 as a *horizontal* shift of the aggregate demand curve in the supply-demand diagram.
4. The supply-demand diagram will now show the actual effect on real output as well as the resulting inflation or deflation.[2]

---

[2]The change in the price level reacts back on the 45° diagram. See Review Question 10 at the end of the chapter.

A ROLE FOR STABILIZATION POLICY

Chapter 25 emphasized the volatility of investment spending, and Chapter 26 noted that changes in investment have multiplier effects on aggregate demand. This chapter took the next step by showing how shifts in the aggregate demand curve cause fluctuations in both real GDP growth and inflation—fluctuations that are widely decried as undesirable. It also suggested that the economy's self-correcting mechanism works, but slowly, thereby leaving room for government stabilization policy to improve the workings of the free market. Can the government really do this? If so, how? These are the questions for the next part of the book.

SUMMARY

1. The economy's **aggregate supply curve** relates the quantity of goods and services that will be supplied to the price level. It normally slopes upward to the right because the costs of labor and other inputs are relatively fixed in the short run, meaning that higher selling prices make input costs relatively "cheaper" and therefore encourage greater production.
2. The position of the aggregate supply curve can be shifted by changes in money wage rates, prices of other inputs, technology, or quantities or qualities of labor and capital.
3. The *equilibrium price level* and the *equilibrium level of real GDP* are jointly determined by the intersection of the economy's aggregate supply and aggregate demand schedules. This intersection may come at full employment, below full employment (a recessionary gap), or above full employment (an inflationary gap).
4. The economy has a self-correcting mechanism that erodes a *recessionary gap*. Specifically, a weak labor market reduces wage increases and, in some cases, may even drive wages down. This shifts the aggregate supply curve outward. But it happens very slowly.
5. If there is an *inflationary gap*, the economy has a similar mechanism that erodes the gap through a process of inflation. Unusually strong job prospects push wages up, which shifts the aggregate supply curve to the left and reduces the inflationary gap.
6. One consequence of this *self-correcting mechanism* is that, if a surge in aggregate demand opens up an inflationary gap, part of the economy's natural adjustment to this event will be a period of stagflation; that is, a period in which prices are rising while output is falling.
7. An inward shift of the aggregate supply curve will cause output to fall while prices rise; that is, it will cause **stagflation.** Among the events that have caused such a shift are abrupt increases in the price of foreign oil.
8. Adverse supply shifts like this plagued our economy when oil prices skyrocketed in 1973 to 1974, 1979 to 1980, and again in 1990, leading to stagflation each time.
9. Among the reasons why the oversimplified multiplier formula is wrong is the fact that it ignores the inflation that is caused by an increase in aggregate demand. Such *inflation decreases the multiplier* by reducing both consumer spending and net exports.

KEY TERMS

Aggregate supply curve
Productivity
Equilibrium of real GDP and the price level
Inflationary gap
Self-correcting mechanism
Stagflation
Recessionary gap
Inflation and the multiplier

QUESTIONS FOR REVIEW

1. In an economy with the following aggregate demand and aggregate supply schedules, find the equilibrium levels of real output and the price level. Graph your solution. If full employment comes at $2,800 billion, is there an inflationary or a recessionary gap?

Aggregate Quantity Demanded (in billions)	Price Level	Aggregate Quantity Supplied (in billions)
3,200	85	2,600
3,100	90	2,750
3,000	95	2,850
2,900	100	2,900
2,800	105	2,925

2. Suppose a worker receives a wage of $18 per hour. Compute the real wage (money wage deflated by the price index) corresponding to each of the following possible price levels: 85, 95, 100, 110, 120. What do you notice about the relationship between the real wage and the price level? Relate this to the slope of the aggregate supply curve.

3. Explain why a decrease in the price of foreign oil shifts the aggregate supply curve outward to the right. What are the consequences of such a shift?

4. Comment on the following statement: "Inflationary and recessionary gaps are nothing to worry about because the economy has a built-in mechanism that cures either type of gap automatically."

5. Give *two* different explanations of how the economy can suffer from stagflation.

6. Why do you think wages tend to be rigid in the downward direction?

7. Add the following aggregate supply and demand schedules to the example in Review Question 3 of Chapter 25 (page 606) to see how inflation affects the multiplier.

   Draw these schedules on a piece of graph paper. Then:

   a. Notice that the difference between Columns 2 and 3 (the aggregate demand schedule at two different levels of investment) is always $200. Discuss how this relates to your answer in the previous chapter.

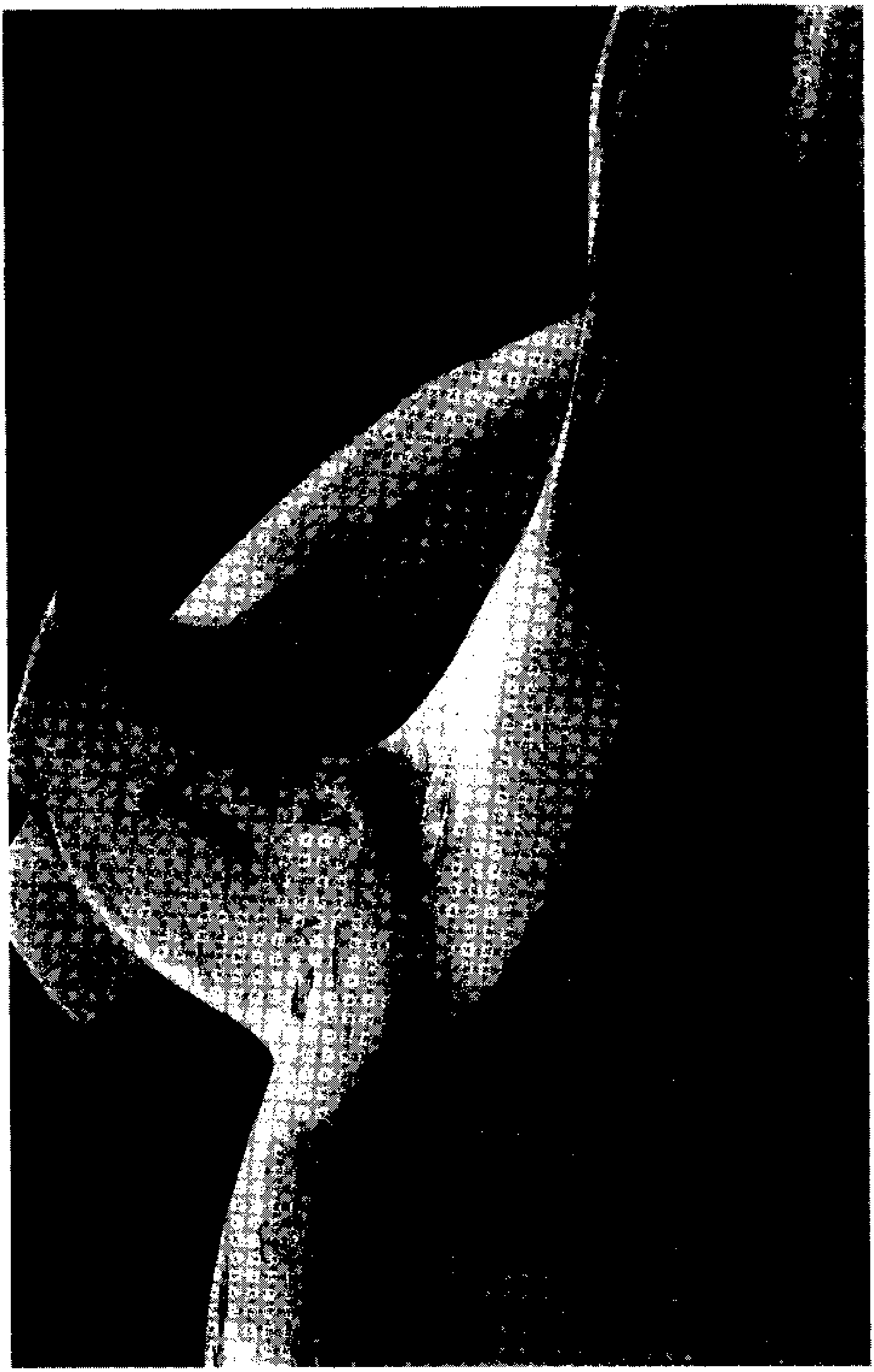
   b. Find the equilibrium GDP and the equilibrium price level both before and after the increase in investment. What is the value of the multiplier?

(1) Price Level	(2) Aggregate Demand (when investment is $240)	(3) Aggregate Demand (when investment is $260)	(4) Aggregate Supply
90	$3,860	$4,060	$3,660
95	3,830	4,030	3,730
100	3,800	4,000	3,800
105	3,770	3,970	3,870
110	3,740	3,940	3,940
115	3,710	3,910	4,010

8. Explain in words why rising prices reduce the multiplier effect of an autonomous increase in aggregate demand.

9. Use an aggregate supply and demand diagram to show that multiplier effects are smaller when the aggregate supply curve is steeper. Which case gives rise to more inflation—the steep aggregate supply curve or the flat one? What happens to the multiplier if the aggregate supply curve is vertical?

10. (More difficult) Assume that investment spending rises. Draw a set of graphs illustrating the Steps in Calculating the Multiplier listed on page 644. Your aggregate supply and demand diagram from Steps 3 and 4 will show a change in the price level. How would this change in the price level react back on the 45° line diagram you used in Step 1? In view of this, use the 45° line diagram to show that inflation reduces the multiplier.



PART VII

Fiscal and Monetary Policy

CHAPTER 28

Managing Aggregate Demand: Fiscal Policy

Next, let us turn to the problems of our fiscal policy. Here the myths are legion and the truth hard to find.
John F. Kennedy

In the model of the economy constructed in Part VI, the government played an entirely passive role. It did a certain amount of spending and collected a fixed amount of taxes, and that was it. We concluded that such an economy has a tendency to move toward an equilibrium with high employment and low inflation, but only a weak one. Furthermore, we hinted that well-designed policies might improve the economy's performance. It is now time to pick up that hint—and to learn some of the difficulties the government must overcome if it is to conduct a successful stabilization policy.

The government's **FISCAL POLICY** is its plan for spending and taxation. It is designed to steer aggregate demand in some desired direction.

Traditionally, the government has used its taxing and spending powers to influence the demand side of the economy. So this chapter begins there, in the domain of **fiscal policy.** The next two chapters take up the government's other main tool for managing aggregate demand: *monetary policy.*

We start by allowing taxes to depend on income—as they do in the real world—and then considering the multipliers for tax policy. As we shall see, this analysis requires no fundamental change in our model of the determination of GDP and the price level. However, it does reduce the size of the multiplier.

As Presidents Reagan, Bush, and Clinton have all realized, the effects of tax policy are not limited to aggregate *demand.* Taxes also affect aggregate *supply.* So the last sections of the chapter examine "supply-side economics."

ISSUE: HOW QUICKLY SHOULD WE BALANCE THE BUDGET?

The debate over reducing the federal budget deficit has been vigorous and often highly partisan. By mid-1995, however, both President Clinton and the Republican-controlled Congress had agreed that the budget should eventually be balanced. Their competing plans differed in many ways, reflecting the divergent priorities of the two parties. But one particularly contentious issue was how long it should take to balance the budget. The Republican plan sought to achieve balance in 7 years; the president's original plan took 10. Who was right?

Both sides agreed that the fundamental reason to reduce the deficit was to raise private investment, and thereby to spur long-run economic growth. We will explore in detail the chain of reasoning in Chapter 32. But, roughly, the point is that budget deficits force the government to borrow money, and the resulting increase in demand for credit raises the cost of borrowing. That makes it more expensive for firms to raise funds to build factories and purchase machinery. Note that the growth tonic works on the economy's supply side: the idea is to augment the supply of capital.

But the road from lower deficits to lower interest rates to higher investment to faster growth is a long one. In the meantime, President Clinton worried, balancing the budget too rapidly might reduce aggregate demand which, as we know, can slow economic growth. So the president originally preferred a 10-year path. But the Republicans were adamant about reaching balance in 7 years, and the president eventually conceded the point.

This chapter will explain the logic underlying the president's concern about aggregate demand. Subsequent chapters will develop the supply-side links and offer some counterarguments. Wait until you finish Chapter 32. Then make up your own mind on this important debate.

INCOME TAXES AND THE CONSUMPTION SCHEDULE

FIXED TAXES are taxes that do not vary with the level of GDP.

VARIABLE TAXES are taxes that do vary with the level of GDP.

To understand how taxes affect equilibrium GDP, we must distinguish between **fixed taxes**, which are the only kind we have considered so far, and **variable taxes**, which are much more important in practice.

Most of the taxes collected by the federal, state, and local governments rise and fall with the level of GDP. In some cases, the reason is obvious: *Personal* and *corporate income tax* collections, for example, depend on how much income there is to be taxed. *Sales tax* receipts depend on GDP because consumer spending is higher when GDP is higher. On the other hand, some types of tax receipts—such as property taxes—do not vary with GDP. We call the first kind of tax *variable taxes* and the second kind *fixed taxes*.

Why is this distinction important? Remember that taxes (T) are the difference between gross domestic product (Y) and disposable income (DI):

$$DI = Y - T$$

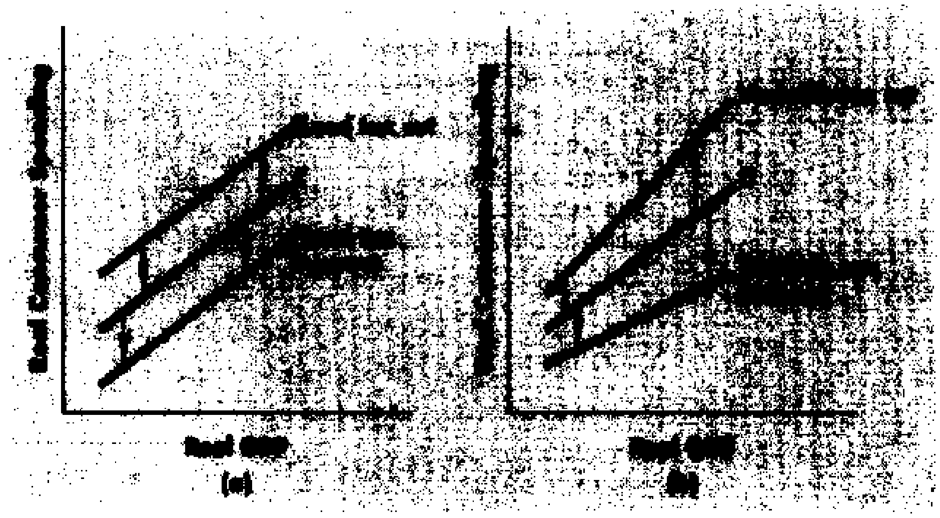
Thus, when taxes are increased, disposable income falls—and hence so does consumption—*even if GDP is unchanged*. As a result:

An increase in taxes shifts the consumption schedule in our 45° line diagram downward. Similarly, a reduction in taxes shifts the consumption schedule upward.

But precisely how the consumption schedule shifts depends on the nature of the tax change. If a fixed tax is increased, disposable income falls by the *same* amount regardless of the level of GDP. Hence, the decline in consumer spending is the same. In a word, the C schedule shifts downward in a parallel manner, as shown in Figure 28-1(a).

But many tax policies change disposable income by amounts that depend on the level of income—normally being larger at higher income levels. This is true, for example, whenever Congress alters the bracket rates in the personal income tax code, as it last did in 1993. Since higher tax rates decrease disposable income more when GDP is higher, the C schedule shifts down more sharply at high income levels than at low ones. [See Figure 28-1(b).]

FIGURE 28-1 HOW TAX POLICY SHIFTS THE CONSUMPTION SCHEDULE



Because consumption depends on disposable income, not GDP, any change in taxes will shift the consumption schedule relating consumption to GDP. Panel (a) shows how the C curve shifts for changes in fixed taxes. Panel (b) shows how the C curve shifts if the tax cut (or tax increase) is larger at high incomes than at low incomes.

The two parts of Figure 28-1 illustrate the first reason why the distinction between fixed and variable taxes is important. Figure 28-2 illustrates the second. Here we show two consumption lines, C_1 and C_2. C_1 is the consumption schedule used in previous chapters; it is constructed on the assumption that taxes are fixed at $1,200 billion regardless of GDP. C_2 depicts a more realistic case in which tax collections are 20 percent of GDP. You will notice that C_2 is flatter than C_1. This is no accident. In fact:

Variable taxes such as the income tax flatten the consumption schedule in a 45° line diagram.

It is easy to understand why. Table 28-1 shows in Column 1 alternative values of GDP ranging from $4.5 trillion to $7.5 trillion. Column 2 then indicates that taxes are always one-fifth of this amount. Column 3 subtracts Column 2 from Column 1 to arrive at disposable income (*DI*). Column 4 then gives the amount of consumer spending corresponding to each level of *DI*. The schedule relating *C* to *Y*, which we need for our 45° line, is therefore found in Columns 1 and 4.

Notice in Table 28-1 that for each $500 billion increase in GDP, consumer spending rises by $300 billion. Thus, the slope of line C_2 in Figure 28-2 is $300/$500, or 0.60. But, if you look back at any of the tables in Chapter 26, you will find that consumption there rose by $300 billion each time GDP increased $400 billion—making the slope $300/$400, or 0.75. (See the steeper line C_1 in Figure 28-2.)

All of this sounds terribly mechanical, but the economic reasoning behind it is straightforward. When taxes are fixed, as in line C_1, each additional dollar of GDP raises disposable income (*DI*) by $1. Consumer spending then rises by $1 times the marginal propensity to consume (MPC), which is 0.75 in our example. Hence, each additional dollar of GDP leads to 75 cents more spending. But when taxes vary with income, each additional dollar of GDP raises *DI* by less than $1 because the government takes a share in taxes. In our example, taxes are 20 percent of GDP, so each additional $1 of GDP leads to 80 cents more *DI*. With an MPC of 0.75, that means that spending rises 60 cents (75 percent of 80 cents) each time GDP rises by $1. That is why the slope of line C_2 in Figure 28-2 is only 0.60, instead of 0.75.

FIGURE 28-2 THE CONSUMPTION SCHEDULE WITH FIXED AND VARIABLE TAXES

Line C_1 is the consumption schedule used in earlier chapters, when taxes were taken to be fixed. If, instead, tax receipts rise with GDP, the consumption schedule is flatter—see line C_2.

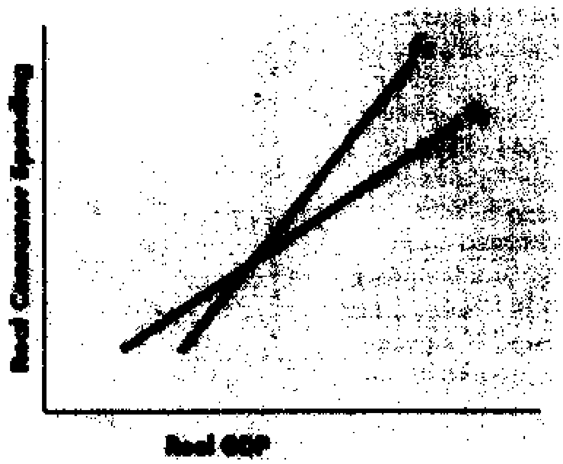


Table 28-2 and Figure 28-3 take the next step by replacing the old consumption schedule with this new one in both the tabular presentation of income determination and the 45° line diagram. We see immediately that the equilibrium level of GDP is at point *E*. Here gross domestic product is $6,000 billion, consumption is $3,900 billion, investment is $900 billion, net exports are −$100 billion, and government purchases are $1,300 billion. As we know, full employment may occur above or below Y = $6,000 billion. If below, there is an inflationary gap. Prices probably will start to rise, pulling the expenditure schedule down and reducing equilibrium GDP. If above, there is a recessionary gap, and history suggests that prices will fall only slowly. In the interim, there will be a period of high unemployment.

TABLE 28-1 THE EFFECTS OF AN INCOME TAX ON THE CONSUMPTION SCHEDULE (BILLIONS OF DOLLARS PER YEAR)

(1) Gross Domestic Product	(2) Taxes	(3) Disposable Income (GDP minus taxes)	(4) Consumption
$4,500	$ 900	$3,600	$3,000
5,000	1,000	4,000	3,300
5,500	1,100	4,400	3,600
6,000	1,200	4,800	3,900
6,500	1,300	5,200	4,200
7,000	1,400	5,600	4,500
7,500	1,500	6,000	4,800

This table shows how an income tax lowers the slope of the consumption schedule (Column 4) in a concrete example. For every $500 billion increase in GDP, consumption rises by $300 billion in this example (compare Columns 1 and 4), so the slope of the C_2 line in Figure 28-2 is $300/$500, or 0.6. In earlier chapters, the slope of the consumption schedule (line C_1 in Figure 28-2) was $300/$400, or 0.75.

TABLE 28-2 TOTAL EXPENDITURE SCHEDULE WITH A 20 PERCENT INCOME TAX (BILLIONS OF DOLLARS PER YEAR)

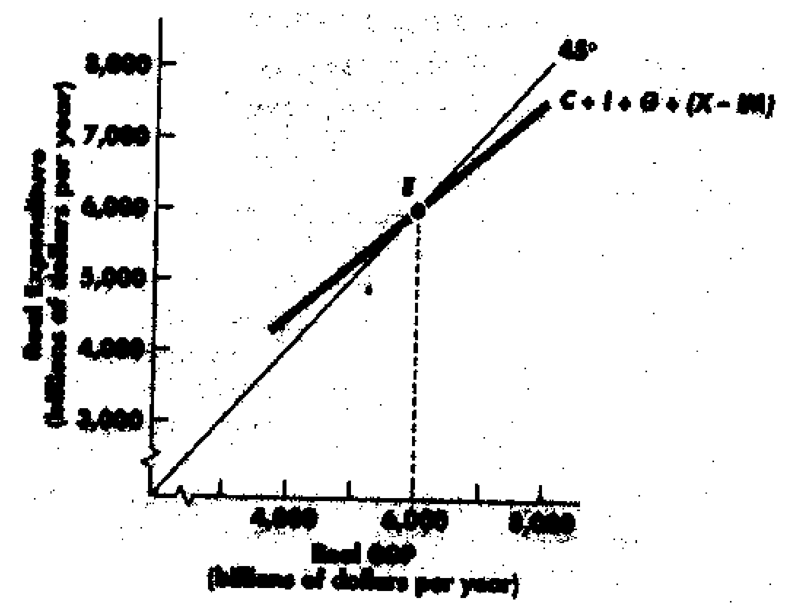
(1) Gross Domestic Product Y	(2) Consumption C	(3) Investment I	(4) Net Exports $(X - IM)$	(5) Government Purchases G	(6) Total Expenditures $C + I + G + (X - IM)$
$4,500	$3,000	$900	−$100	$1,300	$5,100
5,000	3,300	900	−100	1,300	5,400
5,500	3,600	900	−100	1,300	5,700
6,000	3,900	900	−100	1,300	6,000
6,500	4,200	900	−100	1,300	6,300
7,000	4,500	900	−100	1,300	6,600
7,500	4,800	900	−100	1,300	6,900

This table replaces the previous consumption schedule with a new one that adjusts for the income tax (as shown in Table 28–1) and shows that the equilibrium level of income is still $6,000 billion.

In short, once we adjust the expenditure schedule for variable taxes, the determination of national income proceeds exactly as before. The effects of government spending and taxation, therefore, are fairly straightforward and can be summarized as follows:

Government purchases of goods and services add to total spending directly through the G component of $C + I + G + (X - IM)$. Taxes indirectly reduce total spending by lowering disposable income and thus reduce the C component of $C + I + G + (X - IM)$. On balance, then, the government's actions may raise or lower the equilibrium level of GDP, depending on how much spending and taxing it does.

FIGURE 28-3 INCOME DETERMINATION WITH A VARIABLE INCOME TAX



This diagram adds a 20 percent income tax to the model economy portrayed in earlier chapters. Because of this, the $C + I + G + (X - IM)$ schedule is flatter, but otherwise things look just as before. In particular, equilibrium is at point E, where the $C + I + G + (X - IM)$ schedule crosses the 45° line. Thus, equilibrium GDP is $6,000 billion.

However, there is more to the story. As we will see now:

The multiplier is smaller in the presence of an income tax.

THE MULTIPLIER REVISITED

We learned in Chapter 26 that the multiplier works through a chain of spending and respending, as one person's expenditure becomes another's income. But, when there is an income tax, some of the additional income leaks out of the circular flow at each stage. Specifically, if the income tax rate is 20 percent, when Generous Motors spends $1 million on salaries, workers actually receive only $800,000 in *after-tax* (that is, disposable) income. If workers spend 75 percent of this amount (because the MPC is 0.75), spending in the next round will be only $600,000. Notice that this is only *60 percent* of the original expenditure, not *75 percent*—just as we observed in the last section.

Thus, the multiplier chain for each original dollar of spending shrinks from

$$1 + 0.75 + (0.75)^2 + (0.75)^3 + \ldots = \frac{1}{1 - 0.75} = \frac{1}{0.25} = 4$$

to

$$1 + 0.6 + (0.6)^2 + (0.6)^3 + \ldots = \frac{1}{1 - 0.6} = \frac{1}{0.4} = 2.5$$

This is clearly a large reduction in the multiplier. We thus have a third reason why the oversimplified multiplier formula of Chapter 26 exaggerates the size of the multiplier: It ignores income taxes.

REASONS WHY THE OVERSIMPLIFIED MULTIPLIER FORMULA IS WRONG

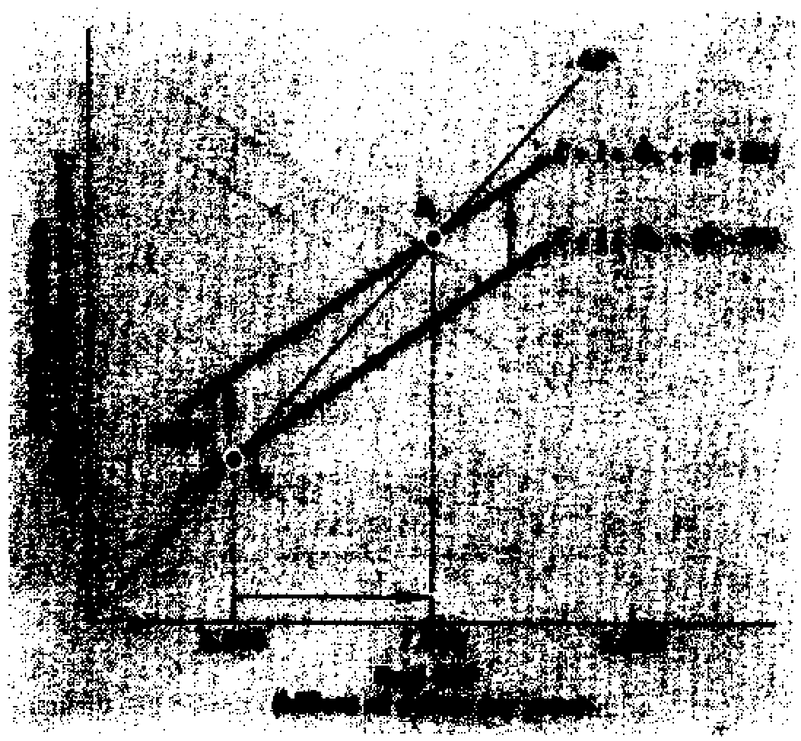
1. It ignores variable imports, which reduce the size of the multiplier.
2. It ignores price-level changes, which reduce the multiplier.
3. It ignores income taxes, which also reduce the size of the multiplier.

The last of these three reasons is very important in practice.

This conclusion about the multiplier is shown graphically in Figure 28-4, where we have drawn our $C + I + G + (X - IM)$ schedules with a slope of 0.6 rather than the 0.75 slope we used in earlier chapters. This new slope reflects an MPC of 0.75 and a tax rate of 20 percent. Figure 28-4 depicts the effect of a $400 billion increase in government purchases of goods and services, which shifts the $C + I + G + (X - IM)$ schedule from $C + I + G_0 + (X - IM)$ to $C + I + G_1 + (X - IM)$. Equilibrium moves from point E_0 to point E_1—an increase of GDP from $Y = \$6,000$ billion to $Y = \$7,000$ billion. Thus, if we ignore for the moment any increases in the price level (which would further reduce the multiplier), a $400 billion increment in government spending leads to a $1,000 billion increment in GDP. So, when taxes are included in our model, the multiplier is only $1,000/$400 = 2.5, just as we concluded before.

FIGURE 28-4 THE MULTIPLIER IN THE PRESENCE OF AN INCOME TAX

This diagram illustrates that an economy with an income tax (in this case a 20 percent income tax) has a lower multiplier than an economy without one. Specifically, the $C + I + G + (X - IM)$ curve is shifted upward by a $400 billion increase in G, and the diagram shows that equilibrium GDP rises by $1,000 billion—from $6,000 billion to $7,000 billion. The multiplier is therefore $1,000/$400 = 2.5, whereas without an income tax it was 4.



MULTIPLIERS FOR TAX POLICY

Because they work indirectly via consumption, multipliers for tax changes are more complicated than multipliers for spending such as G. They must be worked out in two steps.

Step 1. Figure out how much any change in the tax law affects consumer spending.

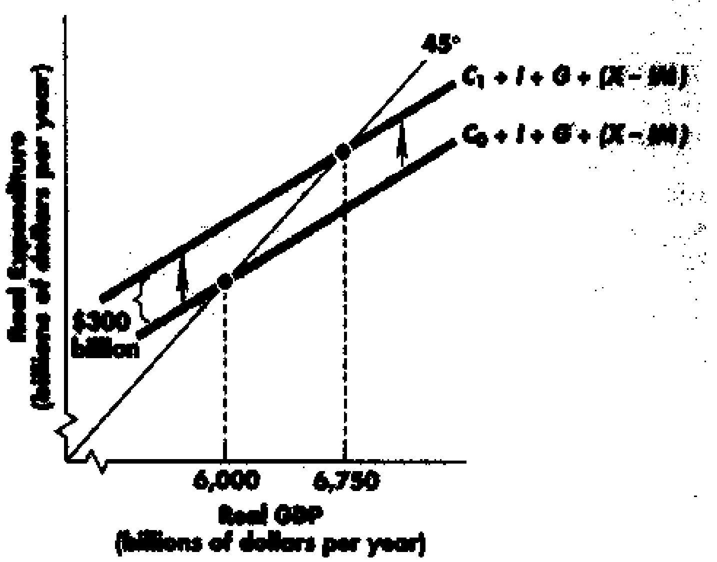
Step 2. Enter this vertical shift of the consumption schedule in the 45° line diagram and see how it affects output.

A reduction in income taxes provides a convenient example of this two-step procedure because we have already done Step 1 in Chapter 24, when we studied how consumer spending would respond to an income tax cut. There, we concluded that, if the tax reduction were viewed as permanent, consumers would increase their spending by an amount equal to the tax cut times the marginal propensity to consume. (If you need review, turn back to pages 568–572.)

To create a simple and familiar numerical example, suppose income taxes fall by $400 billion at each level of GDP—say, from $0.2Y$ (that is, 20 percent of GDP) to $0.2Y$ − $400 billion. Step 1 instructs us to multiply the $400 billion tax cut by the marginal propensity to consume (MPC), which is 0.75, to get $300 billion as the vertical shift of the consumption schedule.

Step 2 then amounts to multiplying this $300 billion increase in consumption by the multiplier—which is 2.5 in our example—giving $750 billion as the rise in GDP. Figure 28-5 verifies that this is so by entering a $300 billion vertical shift

FIGURE 28-5 THE MULTIPLIER FOR A REDUCTION IN FIXED TAXES



In this example, the $C + I + G + (X - IM)$ schedule is shifted vertically upward by $300 billion, from $C_0 + I + G + (X - IM)$ to $C_1 + I + G + (X - IM)$, by a $400 billion tax cut. Equilibrium GDP therefore increases from $6,000 billion to $6,750 billion. So the multiplier is $750/$400 = 1.875.

of the consumption function into the 45° line diagram and noting that GDP does indeed rise by $750 billion as a result.

Notice something interesting here. The $400 billion tax cut raises GDP by $750 billion. Thus, the multiplier is $700/$400 = 1.875. But the multiplier for the $400 government purchases that we worked out on page 654 and depicted in Figure 28-4 was 2.5. What's going on here? Apparently:

The multiplier for changes in taxes is smaller than the multiplier for changes in government purchases.

The reason is not mysterious. While G is a direct component of total expenditure, taxes are not. Taxes work indirectly, by first changing disposable income and then changing C. Since some of the change in disposable income affects *saving* rather than *spending*, a $1 tax cut does not pack as much punch as $1 of G. That is why we must multiply the $400 billion change in taxes by 0.75 to get the $300 billion shift of the C schedule shown in Figure 28-5.

The fact that the multiplier for taxes is smaller than the multiplier for G has a surprising implication:

If government purchases (G) and taxes (T) rise by equal amounts, the government seems to be giving to the public with one hand and taking away an equal amount with the other. But the effects do not cancel out. Instead, the equilibrium level of GDP on the demand side rises. Similarly, if G and T fall by equal amounts, the equilibrium level of GDP on the demand side falls.

For example, we have seen that a $400 billion increase in G raises GDP by $1,000 billion, while a $400 billion hike in T lowers GDP by just $750 billion. Thus, if *both G and T* are raised by $400 billion, GDP will go up by $250 billion.

The moral of the story is that fiscal policies that keep the deficit the same do not necessarily keep aggregate demand the same. A cut in government spending balanced by an equal cut in tax revenues can be expected to reduce Y—a lesson that policymakers frequently forget.

GOVERNMENT TRANSFER PAYMENTS

Finally, we should mention the last major tool of fiscal policy: *government transfer payments*. A transfer, you will remember, is a payment to an individual that is not compensation for any direct contribution to production. How are transfers treated in our models of income determination—like purchases of goods and services (G) or like taxes (T)?

The answer follows readily from the circular flow diagram back on page 593 or the accounting identity on page 568. The important thing to understand about transfer payments is that they intervene between gross domestic product (Y) and disposable income (DI) in precisely the *opposite* way from income taxes.

Specifically, starting with the wages, interest, rents, and profits that constitute the national income, we *subtract* income taxes to calculate disposable income. We do so because these taxes represent the portion of incomes that are *earned* but never *received* by consumers. But then we must *add* transfer payments because they represent sources of income that are *received* although they were not *earned* in the process of production. Thus:

Transfer payments are basically negative taxes.

So giving consumers \$400 billion in the form of transfer payments is treated in the 45° line diagram as a \$400 billion decrease in taxes. Thus, Figure 28-5, which we devised to illustrate a tax cut, can also be used to illustrate a rise in unemployment benefits, or in Social Security benefits, or any other such transfer payment. Similarly, the analysis of a decrease in transfer payments would proceed exactly like the analysis of an increase in taxes.

PLANNING EXPANSIVE FISCAL POLICY

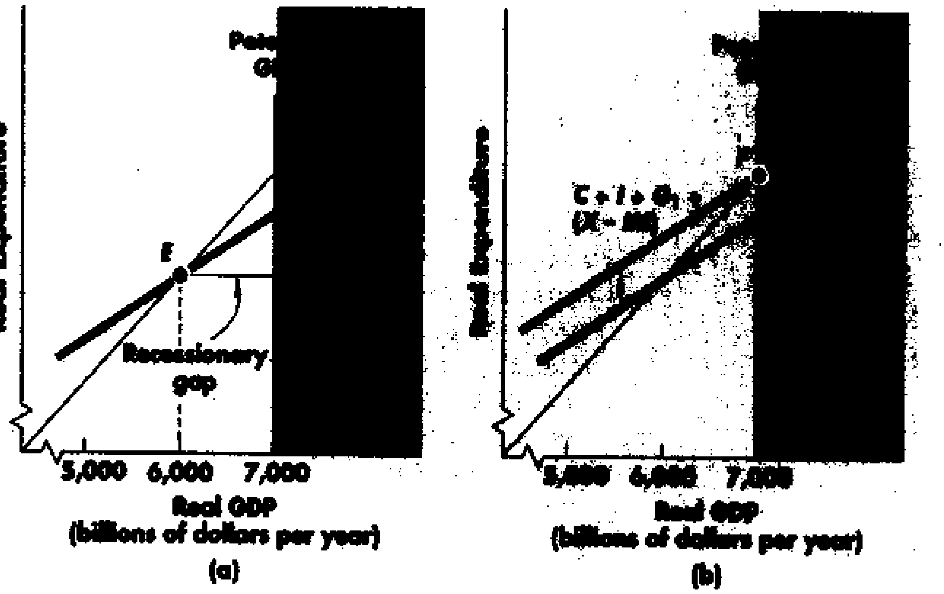
Now imagine you are a member of Congress trying to decide whether to use fiscal policy to stimulate the economy—and, if so, by how much. Suppose the economy would have a GDP of \$6,000 billion if last year's budget were simply reenacted. Suppose further that your goal is to achieve a fully employed labor force and that staff economists tell you this would require a GDP of approximately \$7,000 billion. Finally, just to keep the calculations manageable, imagine that the price level is fixed. What sort of budget should you vote for?

This chapter has taught us that the government has three ways to raise GDP by \$1,000 billion. Congress can raise government purchases, reduce taxes, or increase transfer payments by enough to close the recessionary gap between actual and potential GDP.

Figure 28-6 illustrates the problem, and its cure through higher government spending, on our 45° line diagram. Figure 28-6(a) shows the equilibrium of the economy if no changes are made in the budget. Except for the full-employment line at Y = \$7,000 and the corresponding recessionary gap, it looks just like Figure 28-3. With an expenditure multiplier of 2½, you can figure out that an additional \$400 billion of government spending will be needed to push GDP up \$1,000 billion and eliminate the gap (\$400 × 2½ = \$1,000).

So you might vote to raise G from G_0 = \$1,300 billion to G_1 = \$1,700 billion, hoping to move the $C + I + G + (X - IM)$ line in Figure 28-6(a) up to the position indicated in Figure 28-6(b), thereby achieving full employment. Or you might prefer to achieve this fiscal stimulus by lowering income taxes instead. Or

FIGURE 28-6 FISCAL POLICY TO ELIMINATE A RECESSIONARY GAP



This diagram shows, with more precision than can actually be achieved in practice, how fiscal policy can eliminate a recessionary gap. Panel (a) shows the gap: Equilibrium GDP ($6,000 billion) falls short of potential GDP ($7,000 billion). Panel (b) shows how fiscal policy—by moving the $C + I + G + (X - IM)$ curve up just enough—can wipe out this gap and restore full employment. With a multiplier of 2½, a rise in G of $400 billion or a cut in taxes large enough to shift C up by $400 billion would do the trick.

you might opt for more generous transfer payments. The point is that there are a variety of budgets capable of increasing GDP by $1,000 billion. Figure 28-6 applies equally well to any of them.

PLANNING RESTRICTIVE FISCAL POLICY

The preceding example assumed that the basic problem of fiscal policy is to close a recessionary gap, as was the case at the start of the Clinton administration. Often this is so. But by the end of 1994 many economists believed that the major macroeconomic problem was just the opposite: Aggregate demand exceeded the economy's capacity to produce, leading to an inflationary gap. In such a case, fiscal policy should assume a restrictive stance in order to reduce inflation.

It does not take much imagination to run our previous analysis in reverse. If there is an inflationary gap under a continuation of current budget policies, contractionary fiscal policy tools can eliminate it. By cutting spending, raising taxes, or by some combination of the two, the government can pull the $C + G + I + (X - IM)$ schedule down to a noninflationary position and achieve an equilibrium at full employment.

Notice the difference between this way of eliminating an inflationary gap and the natural self-correcting mechanism that we discussed in the last chapter. There we observed that, if the economy were left to its own devices, a cumulative but self-limiting process of inflation eventually would eliminate the inflationary gap and return the economy to full employment. Here we see that it is not necessary to put the economy through the inflationary wringer. Instead, a restrictive fiscal policy can avoid inflation by limiting aggregate demand to the level the economy can produce at full employment.

But suppose taxes are raised or spending is cut when the economy is approximately at full employment, rather than stuck with an inflationary gap. Then contractionary fiscal policy can cause unemployment. This is the problem that worried President Clinton in 1995, when he argued at first that Republican plans to balance the budget in 7 years were too much, too soon.

THE CHOICE BETWEEN SPENDING POLICY AND TAX POLICY

In principle, fiscal policy can nudge the economy in the desired direction equally well by changing government spending or by changing taxes. For example, if the government wants to expand the economy, it can raise G or lower T. Either policy would shift the total expenditure schedule upward, as depicted in Figure 28-6, thereby raising the equilibrium GDP on the demand side.

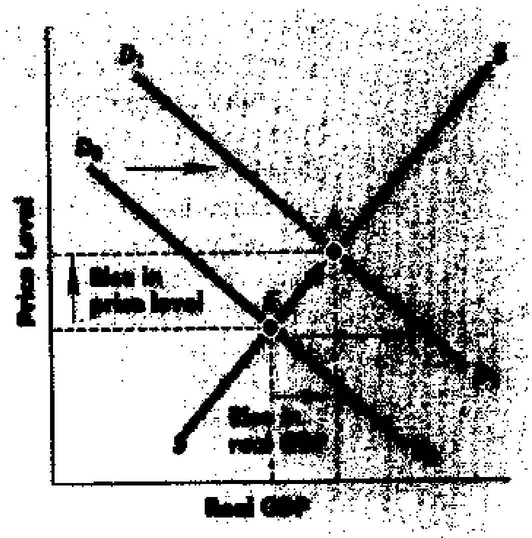
In terms of our aggregate demand and supply diagram, either policy shifts the aggregate demand curve outward, from D_0D_0 to D_1D_1 in Figure 28-7. As a result, the economy's equilibrium moves from point E to point A. Both real GDP and the price level rise. As this diagram points out, any combination of higher spending and lower taxes that produces the same aggregate demand curve leads to the same increases in real GDP and prices.

How, then, do policymakers decide whether it is better to raise spending or to cut taxes? The answer depends mainly on how large a public sector they want—a long-running debate in the United States which has become particularly contentious of late.

The small-government point of view is typically advocated by conservatives, who gained the ascendancy in the 1990s. According to this view, the government is much too large; we are foolish to rely on the public sector to do things that private individuals and businesses could do better on their own; and the growth of government interferes too much in our everyday lives, and in so doing circumscribes our freedom. Those who hold this view can argue for *tax cuts* when macroeconomic considerations call for expansionary fiscal policy, and for *reductions in public spending* when restrictive policy is required.

An opposing opinion, expressed most often by liberals, is that there is something amiss when a country as wealthy as the United States has such an impoverished public sector. In this view, America's most pressing needs are not for more fast food and video games, but rather for better schools, more efficient public transportation systems, and cleaner and safer city streets. People on this side of the debate believe that we should *increase* spending when the economy needs

FIGURE 28-7 EXPANSIONARY FISCAL POLICY



Any of a variety of expansionary fiscal policies will push the aggregate demand curve outward to the right as depicted by the shift from D_0D_0 to D_1D_1 in this aggregate supply and demand diagram. The economy's equilibrium moves upward to the right along aggregate supply curve SS, from point E to point A. Comparing A with E, we note that output is higher but prices are also higher. The expansionary policy has caused some inflation.

stimulus, and pay for these improved public services by *increasing taxes* when the economy needs to be reined in.

Too often the use of fiscal policy for economic stabilization is erroneously associated with a large and growing public sector—that is, with "big government." This need not be the case. Individuals favoring a smaller public sector can advocate an active fiscal policy just as well as those who favor a larger public sector. Advocates of big government budgets should seek to expand demand (when appropriate) through higher government spending and contract demand (when appropriate) through tax increases. By contrast, advocates of small public budgets should seek to expand demand by cutting taxes and reduce demand by cutting expenditures.

SOME HARSH REALITIES

The mechanics outlined so far in this chapter make the fiscal policy planner's job look rather simple. The elementary diagrams suggest, rather misleadingly, that the authorities can drive GDP to any level they please simply by manipulating spending and tax programs. It seems as though they should be able to hit the full-employment bull's eye every time.

But, in fact, a better analogy is shooting through dense fog at an erratically moving target with an inaccurate gun. The target is moving because, in the real world, the investment, net exports, and consumption schedules are constantly shifting due to changes in expectations, new technological breakthroughs, events abroad, and so on. This means that the policies decided upon today, which are to take effect at some future date, may no longer be appropriate by the time that date rolls around. Policy must be based, to some extent, on *forecasting*, and no one has yet discovered a foolproof method of economic forecasting.[1] Because fiscal policy decisions sometimes take a long time to be carried out, poor forecasts may occasionally leave the government fighting the last inflation just when the new recession gets under way.

A second misleading feature of our diagrams is that multipliers are not known with as much precision as our examples suggest. Thus, while the "best guess" may be that a $20 billion cut in government purchases will reduce GDP by $40 billion, the actual outcome may be as little as $20 billion or as much as $60 billion. It is therefore impossible to "fine tune" every wobble out of the economy's growth path; economic science is simply not that precise.

A third complication is that our target—full-employment GDP—may be only dimly visible, as if through a fog. For example, there has recently been a lively debate over whether mainstream economists have overestimated the unemployment rate that corresponds to full employment.

Finally, in trying to decide whether to push the unemployment rate lower, legislators would like to know how large the inflation cost is likely to be. As we know, an expansionary fiscal policy that reduces a recessionary gap by increasing aggregate demand will lower unemployment. But it also tends to be inflationary. This undesirable side effect may make the government hesitant to use fiscal policy to combat recessions.

[1]Some problems and techniques of economic forecasting are considered in Chapter 31.

Is there a way out of this dilemma? Can we carry on the battle against unemployment without aggravating inflation? For almost 20 years, a small but influential minority of economists, journalists, and politicians have argued that we can. They call their approach "supply-side economics." The idea helped sweep Ronald Reagan to smashing electoral victories in 1980 and 1984. By 1992, however, Bill Clinton won the presidency by running against it. Then, just 2 years later, supply-side thinking made something of a comeback when the Republicans swept control of both houses of Congress and began trying to implement their "Contract with America." Just what is supply-side economics?

THE IDEA BEHIND SUPPLY-SIDE TAX CUTS

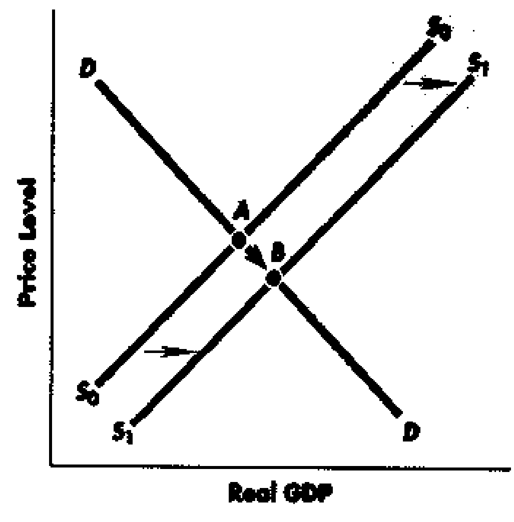
The central idea of supply-side economics is that certain types of tax cuts can be expected to increase aggregate supply. For example, taxes can be cut in ways that raise the rewards for working, saving, and investing. *If people actually respond to these incentives,* such tax cuts will increase the total supplies of labor and capital in the economy, thereby increasing aggregate supply.

Figure 28-8 illustrates the idea on an aggregate supply and demand diagram. If policy measures can shift the economy's aggregate supply to position S_1S_1, then prices will be lower and output higher than if the aggregate supply curve were S_0S_0. Policymakers will have succeeded in reducing inflation and raising real output (lowering unemployment) at the same time. The trade-off between inflation and unemployment will have been defeated. This is the goal of supply-side economics.

What sorts of policies do supply siders advocate? Here is a small sample of their long list of recommended tax cuts:

1. *Lower personal income tax rates.* Sharp cuts in personal taxes were the cornerstone of President Reagan's economic strategy. Tax rates on individuals were

FIGURE 28-8 THE GOAL OF SUPPLY-SIDE TAX CUTS



The basic goal of supply-side tax cuts is to shift the economy's aggregate supply curve outward to the right. For example, the aggregate supply curve might be S_1S_1 under a program of supply-side tax cuts, whereas it would only be S_0S_0 without such tax cuts. If aggregate demand is the same in either case, the supply-side tax cuts would lead to the equilibrium point *B* instead of the equilibrium point *A*. Comparing *B* with *A*, we see that the program reduces prices and raises output.

reduced in stages between 1981 and 1984 and then again in 1986. By 1987, the richest Americans were in the 28 percent tax bracket, and most taxpayers were in the 15 percent bracket. Such low tax rates, supply siders argued, augment the supplies of both labor and capital.

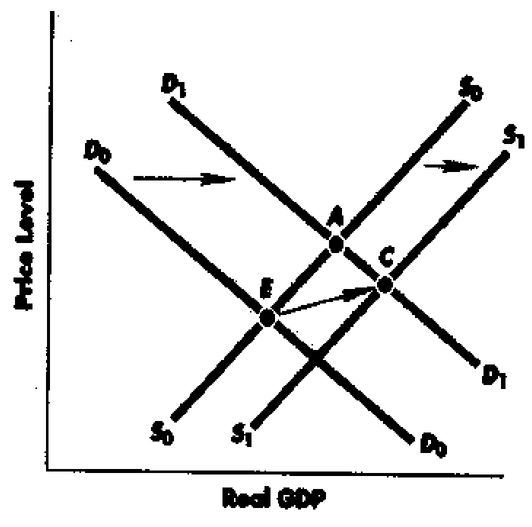
2. *Reduce taxes on income from savings.* One extreme form of this proposal would simply exempt from taxation all income from interest and dividends. Since income must be either consumed or saved, this would, in effect, change our present personal income tax into a tax on consumer spending. Several such proposals for radical tax reform were being seriously considered as this edition went to press.
3. *Reduce taxes on capital gains.* When investors sell assets for a profit, that profit is called a *capital gain* on their investments. Supply siders argue that the government can encourage more investment by taxing capital gains at lower rates than ordinary income. This proposal, long favored by many Republicans, was part of the Contract with America in 1994.
4. *Reduce the corporate income tax.* By reducing the tax burden on corporations, it is argued, the government can provide both greater investment incentives (by raising the profitability of investment) and more investable funds (by letting companies keep more of their earnings). This advice was followed in 1981 by making depreciation allowances more generous[2] and in 1986 by lowering the corporate tax rate.

Let us suppose, for the moment, that a successful supply-side tax cut is enacted. Since *both* aggregate demand *and* aggregate supply increase simultaneously, the economy may be able to avoid the painful inflationary consequences of an expansionary fiscal policy that were shown in Figure 28-7 on page 659.

Figure 28-9 illustrates this conclusion. The two aggregate demand curves and the initial aggregate supply curve S_0S_0 are carried over directly from Figure

FIGURE 28-9 A SUCCESSFUL SUPPLY-SIDE TAX REDUCTION

A supply-side tax cut, if successful, shifts *both* aggregate demand *and* supply to the right. Equilibrium is initially at point E, where demand curve D_0D_0 intersects supply curve S_0S_0. After the supply-side tax cut, the aggregate demand and supply curves are D_1D_1 and S_1S_1, so equilibrium is at point C. Compared to a tax cut that works only on the demand side (point A), the supply-side tax cut raises output more and prices less.



[2]A company investing in a machine or factory may not deduct the entire cost of that asset as a business expense in the year it is purchased. Instead, it must spread the cost over the lifetime of the asset in a series of tax deductions called *depreciation allowances*. Naturally, firms prefer to get these allowances sooner rather than later, because higher depreciation in the early years of an investment means an immediate cut in tax burdens.

28-7. But we have introduced an additional supply curve, S_1S_1, to reflect the successful supply-side tax cut depicted in Figure 28-8. The equilibrium of the economy moves from E to C, whereas with a conventional demand-side tax cut it would have moved from E to A. As compared with point A, output is higher and prices are lower at point C.

A good deal, you say! Indeed it is. The supply-side argument is extremely attractive in principle. The question is: Does it work in practice? Can we actually do what is depicted in Figure 28-9? Let us consider some difficulties.

SOME FLIES IN THE OINTMENT

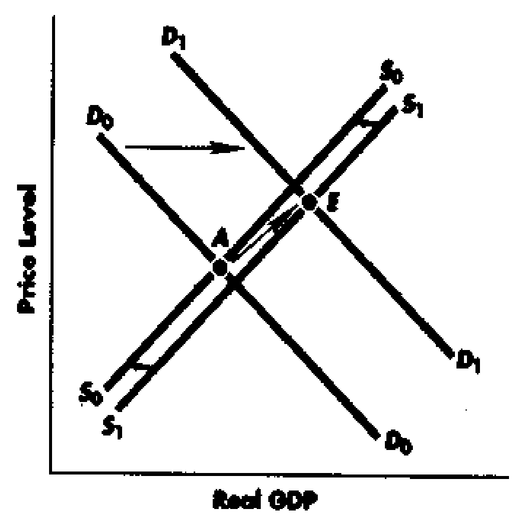
Critics of supply-side economics rarely question its goals or the basic idea that lower taxes improve incentives. They argue, instead, that supply siders exaggerate the beneficial effects of tax cuts and ignore some undesirable side effects. Here is a brief rundown of some of the main objections to supply-side tax cuts:

1. *Small magnitude of supply-side effects.* The first objection is that supply siders are simply too optimistic: We really do not know how to do what Figure 28-8 shows. While it is easy, for example, to design tax incentives that make saving more *attractive* financially, people may not actually respond to these incentives. In fact, most of the statistical evidence suggests that we should not expect very much from tax incentives for saving. As economist Charles Schultze once quipped: "There's nothing wrong with supply-side economics that division by 10 couldn't cure."
2. *Demand-side effects.* The second objection is that supply siders underestimate the effects of tax cuts on aggregate demand. If you cut personal taxes, for example, individuals *may possibly* work more. But they *will certainly* spend more. If you reduce business taxes and thereby encourage expansion of industrial capacity, business firms will demand more investment goods.

   The joint implication of these two objections is shown in Figure 28-10. Here we depict a small outward shift of the aggregate supply curve (which reflects the first objection) and a large outward shift of the aggregate demand curve

FIGURE 28-10 A MORE PESSIMISTIC VIEW OF SUPPLY-SIDE TAX CUTS

If the effect of supply-side tax initiatives on the aggregate supply curve is actually much smaller than suggested by Figure 28-8, the anti-inflationary impact will be correspondingly smaller. As you can see in this diagram, it is possible that a large shift in the aggregate demand curve could overwhelm the favorable effects of the tax cuts on the price level.



(which reflects the second). The result is that the economy's equilibrium moves from point A (the intersection of S_0S_0 and D_0D_0) to point E (the intersection of S_1S_1 and D_1D_1). Prices rise as output expands. The outcome differs only a little from the straight "demand-side" fiscal stimulus depicted in Figure 28-7 (page 659).

3. *Problems in timing.* The most promising types of supply-side tax cuts seek to encourage greater business investment. But the benefits from investment do not arrive by overnight mail. It may take years before we see any substantial increases in industrial capacity. Thus, the *expenditures* on investment goods almost certainly come before the *expansion of capacity.* In a word, supply-side tax cuts will have their primary short-run effects on aggregate demand. Effects on aggregate supply come later.
4. *Effect on the distribution of income.* The preceding objections all pertain to the likely effects of supply-side policies on aggregate supply and demand. But there is a different problem that bears mention: Most supply-side initiatives increase income inequality. While raising the incomes of the wealthiest members of our society may not be their primary aim, most supply-side cuts cannot help but concentrate their benefits on the rich simply because it is the rich who own most of the capital.

   Indeed, this tilt toward the rich is almost an inescapable corollary of supply-side logic. The basic aim of supply-side economics is to increase the incentives for working and investing, that is, to increase the gap between the rewards of those who succeed in the economic game (by working hard, investing well, or by just plain luck) and those who fail. It can hardly be surprising, therefore, that supply-side policies tend to increase economic inequality.
5. *Losses of tax revenue.* You can hardly help noticing that most of the policies suggested by supply siders involve cutting one tax or another. Thus, unless some other tax is raised or spending is cut, supply-side tax cuts are bound to raise the government budget deficit. This problem proved to be the Achilles heel of supply-side economics in the United States in the 1980s. It left behind a legacy of budget deficits that have plagued us ever since.

TOWARD ASSESSMENT OF SUPPLY-SIDE ECONOMICS

On balance, most economists have reached the following conclusions about supply-side tax initiatives:

1. The likely effectiveness of supply-side tax cuts depends on what kinds of taxes are cut. Tax reductions aimed at stimulating business investment are likely to pack more punch than tax reductions aimed at getting people to work longer hours or to save more.
2. Such tax cuts probably will increase aggregate supply much more slowly than they increase aggregate demand. Thus, supply-side policies should not be thought of as a substitute for short-run stabilization policy, but rather as a way to promote (slightly) faster economic growth in the long run.
3. Demand-side effects of supply-side tax cuts are likely to overwhelm supply-side effects in the short run.
4. Supply-side tax cuts are likely to widen income inequalities.
5. Supply-side tax cuts are almost certain to lead to bigger, not smaller, budget deficits.

But this list does not close the books on the issue. It does not even tell us whether supply-side tax cuts are a good idea or a bad one. Some people will look over this list and decide that they favor supply-side tax cuts; others, perusing the same facts, will reach the opposite conclusion. We cannot say that either group is "wrong" because, like almost every economic policy, supply-side economics has its pros and cons.

Why, then, did so many economists and politicians react so negatively to supply-side economics as practiced in the early 1980s? The main reason seems to be that the claims made by the most ardent supply siders were clearly excessive. Naturally, these claims were proven wrong. But showing that wild claims are wild does not dispose of the kernel of truth in supply-side economics: Reductions in marginal tax rates do improve economic incentives. Any specific supply-side tax cut must be judged on its individual merits.

CLINTONOMICS AND SUPPLY-SIDE ECONOMICS

During the 1992 presidential campaign, candidate Bill Clinton attacked the Reagan-Bush brand of supply-side economics as "trickle-down economics," and argued that it had failed. He emphasized, in particular, the last two items in our assessment: the effects on income inequality and on the budget deficit. The voters apparently agreed.

Ironically, however, President Clinton also ran on an avowedly supply-side platform—though of a different stripe. His emphasis was on building up the nation's resources of labor, capital, and technology so that our capacity to produce will be higher in the long run. This, of course, is precisely the goal of supply-side economics: to push the aggregate supply curve outward.

But, unlike Presidents Reagan and Bush, President Clinton did not propose to accomplish this mainly by cutting taxes. While some tax cuts were included in the Clinton plan, especially tax credits for investment and R&D, there were more tax *increases* than tax cuts—in order to raise revenue. The real emphasis of Clintonomics was on improving the quality of the American work force through more and better education and training, and on building up the nation's infrastructure.

But programs like these take money, money the federal government did not have because of the huge budget deficit. So it turned out that the focus of Clintonomics in practice was on deficit reduction, not on public spending designed to raise aggregate supply. In 1993, at a time when Democrats held a majority in both houses of Congress, the president barely won passage of a large deficit-reduction passage. Then, in November 1994, the Republicans won a resounding electoral victory and the political debate tilted even more toward shrinking both the budget deficit and the size of the federal government. How to balance the budget became the central political question of 1995 and 1996. This focus left little room for supply-side initiatives, whether in the form of tax cuts or new spending.

SUMMARY

1. The government's **fiscal policy** is its plan for managing aggregate demand through its spending and taxing programs. It is made jointly by the president and Congress.
2. The government's net effect on aggregate demand—and hence on equilibrium output and prices—depends on whether the expansionary effects of its spending are greater or smaller than the contractionary effects of its taxes.

3. Since consumer spending (*C*) depends on disposable income (*DI*), and *DI* is GDP minus taxes, any change in taxes will shift the consumption schedule on a 45° line diagram. The nature of this shift depends on whether it is **fixed taxes** or **variable taxes** that are changed.
4. Such shifts in the consumption function caused by tax policy are subject to the same multiplier as autonomous shifts in *G*, *I*, or *X* − *IM*.
5. An income tax reduces the size of this common multiplier.
6. The multiplier for changes in taxes is smaller than the multiplier for changes in government purchases.
7. *Government transfer payments* are like negative taxes, not like government purchases of goods and services, because they influence total spending only indirectly through their effect on consumption.
8. If the multipliers were known precisely, it would be possible to plan any of a variety of fiscal policies to eliminate either a recessionary or an inflationary gap. Recessionary gaps can be cured by raising *G*, cutting taxes, or increasing transfers. Inflationary gaps can be cured by cutting *G*, raising taxes, or reducing transfers.
9. Active stabilization policy can be carried out either by means that tend to expand the size of government (by raising either *G* or *T* when appropriate) or by means that hold back the size of government (by reducing either *G* or *T* when appropriate).
10. Expansionary fiscal policy can cure recessions, but it normally exacts a cost in terms of higher inflation. This dilemma has led to a great deal of interest in *"supply-side" tax cuts* designed to stimulate aggregate supply.
11. Supply-side tax cuts aim to push the economy's aggregate supply curve outward to the right. If successful, they can expand the economy and reduce inflation at the same time—a highly desirable outcome.
12. But critics point out at least five serious problems with supply-side tax cuts: They also stimulate aggregate demand; the beneficial effects on aggregate supply may be small; the demand-side effects occur before the supply-side effects; they make the income distribution more unequal; and large tax cuts lead to large budget deficits.

KEY TERMS

Fiscal policy
Fixed taxes
Variable taxes
Government transfer payments
Effect of income taxes on the multiplier
Supply-side tax cuts

QUESTIONS FOR REVIEW

1. America's defense budget has fallen since the end of the Cold War. How would GDP in the United States be affected if the reduced defense spending were used to:
   a. Reduce the budget deficit, so that government purchases fell?
   b. Free funds for other public purposes, so that government purchases remained the same?
2. Consider an economy in which tax collections are always $800 and in which the four components of aggregate demand are as follows:

GDP	Taxes	DI	C	I	G	(X − IM)
$1360	$800	$560	$420	$200	$800	$30
1480	800	680	510	200	800	30
1600	800	800	600	200	800	30
1720	800	920	690	200	800	30
1840	800	1040	780	200	800	30

   Find the equilibrium of this economy graphically. What is the marginal propensity to consume? What is the multiplier? What would happen to equilibrium GDP if government purchases were reduced by $60 and the price level were unchanged?
3. Now consider a related economy in which investment is also $200, government purchases are also $800, net exports are also $30, and the price level is also fixed. But taxes now vary with income, and as a result the consumption schedule looks like the following:

GDP	Taxes	DI	C
$1360	$720	$640	$510
1480	760	720	570
1600	800	800	630
1720	840	880	690
1840	880	960	750

Find the equilibrium graphically. What is the marginal propensity to consume? What is the tax rate? Use your diagram to show the effect of a decrease of \$60 in government purchases. What is the multiplier? Compare this answer to your answer to Review Question 2 above. What do you conclude?

4. Explain why G has the same multiplier as I, but taxes have a different multiplier.
5. Return to the hypothetical economy in Review Question 2 and suppose that *both* taxes and government purchases are increased by \$120. Find the new equilibrium under the assumption that consumer spending continues to be exactly three-quarters of disposable income (as it is in Review Question 2).
6. If the government today decides that aggregate demand is excessive and is causing inflation, what options are open to it? What if it decides that aggregate demand is too weak instead?
7. Suppose that you are in charge of the fiscal policy of the economy in Review Question 2. There is an inflationary gap and you want to reduce income by \$120. What specific actions can you take to achieve this goal?
8. Now put yourself in charge of the economy in Review Question 3, and suppose that full employment comes at a GDP of \$1,840. How can you push income up to that level?
9. Which of the proposed supply-side tax cuts appeals to you most? Draw up a list of arguments for and against enacting such a cut right now.
10. (More difficult) Advocates of lower taxes on capital gains argue that this type of tax cut will raise aggregate supply by spurring business investment. But, of course, any increase in investment spending will also raise aggregate demand. Compare the effects on aggregate supply, aggregate demand, and tax revenue of three different ways to cut the capital gains tax:
    a. Reduce capital gains taxes on *all* investments, including those that were made before tax rates were cut.
    b. Reduce capital gains taxes only on investments made after tax rates are cut.
    c. Reduce capital gains taxes only on certain types of investments, such as corporate stocks and bonds.

    Which of the three seems most desirable to you? Why?

APPENDIX

ALGEBRAIC TREATMENT OF FISCAL POLICY AND AGGREGATE DEMAND

In this appendix, we explain the simple algebra behind the fiscal policy multipliers discussed in the chapter. In so doing, we deal only with a simplified case in which prices do not change. While it is possible to work out the corresponding algebra for the more realistic aggregate demand–aggregate supply analysis with variable prices, the analysis is rather complicated and is best left to more advanced courses.

We start with the example used in the chapter (especially on pages 653 and 655). The government spends \$1,300 billion on goods and services ($G = 1{,}300$) and levies an income tax equal to 20 percent of GDP. So, if the symbol T denotes tax receipts:

$$T = 0.20Y$$

Since the consumption function we have been working with is:

$$C = 300 + 0.75DI$$

where DI is disposable income, and since disposable income and GDP are related by the accounting identity:

$$DI = Y - T$$

it follows that the C schedule used in the 45° line diagram is described by the algebraic equation:

$$\begin{aligned} C &= 300 + 0.75(Y - T) \\ &= 300 + 0.75(Y - 0.20Y) \\ &= 300 + 0.75(0.80Y) \\ &= 300 + 0.60Y \end{aligned}$$

We can now apply the equilibrium condition:

$$Y = C + I + G + (X - IM)$$

Since investment in this example is $I = 900$ and net exports are -100, substituting for C, I, G, and $(X - IM)$ into this equation gives:

$$\begin{aligned} Y &= 300 + 0.60Y + 900 + 1{,}300 - 100 \\ 0.40Y &= 2{,}400 \\ Y &= 6{,}000 \end{aligned}$$

This is all there is to finding equilibrium GDP in an economy with a government.

To find the multiplier for government spending, increase G by 1 and solve the problem again:

$$\begin{aligned} Y &= C + I + G + (X - IM) \\ Y &= 300 + 0.60Y + 900 + 1{,}301 - 100 \\ 0.40Y &= 2{,}401 \\ Y &= 6{,}002.5 \end{aligned}$$

So the multiplier is 6,002.5 − 6,000 = 2.5, as stated in the text.

To find the multiplier for an increase in fixed taxes, change the tax schedule to:

$$T = 0.20Y + 1$$

Disposable income is then:

$$\begin{aligned} DI &= Y - T = Y - (0.20Y + 1) \\ &= 0.80Y - 1 \end{aligned}$$

so the consumption function is:

$$\begin{aligned} C &= 300 + 0.75DI \\ &= 300 + 0.75(0.80Y - 1) \\ &= 299.25 + 0.60Y \end{aligned}$$

Solving for equilibrium GDP as usual gives:

$$\begin{aligned} Y &= C + I + G + (X - IM) \\ Y &= 299.25 + 0.60Y + 900 + 1{,}300 - 100 \\ 0.40Y &= 2{,}399.25 \\ Y &= 5{,}998.125 \end{aligned}$$

So a \$1 increase in fixed taxes lowers Y by \$1.875. The tax multiplier is −1.875.

Now let us proceed to a more general solution, using symbols rather than specific numbers. The equations of the model are as follows:

$$Y = C + I + G + (X - IM) \tag{1}$$

is the usual equilibrium condition;

$$C = a + bDI \tag{2}$$

is the same consumption function we have used in the appendixes of Chapters 25 and 26;

$$DI = Y - T \tag{3}$$

is the accounting identity relating disposable income to GDP;

$$T = T_0 + tY \tag{4}$$

is the tax function, where T_0 represents fixed taxes (which are zero in our numerical example) and t represents the tax rate (which is 0.20 in the example). Finally, I, G, and $(X - IM)$ are just fixed numbers.

We begin the solution by substituting (3) and (4) into (2) to derive the consumption schedule relating C to Y:

$$\begin{aligned} C &= a + bDI \\ C &= a + b(Y - T) \\ C &= a + b(Y - T_0 - tY) \\ C &= a - bT_0 + b(1 - t)Y \end{aligned} \tag{5}$$

You will notice that a change in fixed taxes (T_0) shifts the *intercept* of the C schedule while a change in the tax rate (t) changes its *slope*, as explained in the chapter (pages 650–654).

Next substitute Equation (5) into Equation (1) to find equilibrium GDP:

$$\begin{aligned} Y &= C + I + G + (X - IM) \\ Y &= a - bT_0 + b(1 - t)Y + I + G + (X - IM) \\ [1 - b(1 - t)]Y &= a - bT_0 + I + G + (X - IM) \end{aligned}$$

or

$$Y = \frac{a - bT_0 + I + G + (X - IM)}{1 - b(1 - t)} \tag{6}$$

Equation (6) shows us that the multiplier for G, I, a, or $(X - IM)$ is:

$$\text{Multiplier} = \frac{1}{1 - b(1 - t)}$$

To see that this is in fact the multiplier, raise any of G, I, a, or $(X - IM)$ by 1 unit. In each case, Equation (6) would be changed to read:

$$Y = \frac{a - bT_0 + I + G + (X - IM) + 1}{1 - b(1 - t)}$$

Subtracting Equation (6) from this expression gives the change in Y stemming from a 1-unit change in G or I or a:

$$\text{Change in } Y = \frac{1}{1 - b(1 - t)}$$

We noted in Chapter 26 (page 615) that if there were no income tax ($t = 0$), a realistic value for b (the marginal propensity to consume) would yield a multiplier of 10, which is much bigger than the true multiplier. Now that we have added taxes to the model, our multiplier formula produces much more realistic numbers. Approximate values for these parameters for the U.S. economy are $b = 0.90$ and $t = \frac{1}{3}$. The multiplier formula then gives:

$$\text{Multiplier} = \frac{1}{1 - 0.90(1 - \frac{1}{3})}$$
$$= \frac{1}{1 - \frac{3}{5}} = \frac{1}{\frac{2}{5}}$$
$$= 2.50$$

which is not far from its actual estimated value, nearly 2.

Finally, we can see from Equation (6) that the multiplier for a change in fixed taxes (T_0) is:

$$\text{Tax multiplier} = \frac{-b}{1 - b(1 - t)}$$

For the example considered in the text and earlier in this appendix, $b = 0.75$ and $t = 0.20$, so the formula gives:

$$\frac{-0.75}{1 - 0.75(1 - 0.20)} = \frac{-0.75}{1 - 0.75(0.80)}$$
$$\frac{-0.75}{1 - 0.60} = \frac{-0.75}{0.40} = -1.875$$

According to these figures, each $1 *increase* in T_0 *reduces* Y by $1.875.

QUESTIONS FOR REVIEW

1. Consider an economy described by the following set of equations:

$$C = 120 + 0.80DI$$
$$I = 320$$
$$G = 480$$
$$(X - IM) = -80$$
$$T = 200 + 0.25Y$$

Find the equilibrium level of GDP. Then find the multipliers for government purchases and for fixed taxes. If full employment comes at $Y = 1{,}800$, what are some policies that would get GDP there?

2. This is a variant of the previous problem that approaches things the way a fiscal policy planner might. In an economy whose consumption function and tax function are as given in Review Question 1, with investment fixed at 320 and net exports fixed at −80, find the value of G that would make GDP equal to 1,800.

3. You are given the following information about an economy.

$$C = 0.90DI$$
$$I = 100$$
$$G = 540$$
$$(X - IM) = -40$$
$$T = (\tfrac{1}{3})Y$$

   a. Find equilibrium GDP and the budget deficit.
   b. Suppose the government, unhappy with the budget deficit, decides to cut government spending by precisely the amount of the deficit you found in Review Question 3(a). What actually happens to GDP and the budget deficit, and why?

4. (More difficult) In the economy considered in Review Question 3, suppose the government, seeing that it has not wiped out the deficit, keeps cutting G until it succeeds in balancing the budget. What levels of GDP will then prevail?

CHAPTER 29

MONEY AND THE BANKING SYSTEM

[Money] is a machine for doing quickly and commodiously what would be done, though less quickly and commodiously, without it.
John Stuart Mill

The circular flow diagrams of earlier chapters had a "financial system" in the upper left-hand corner. Savings flowed into this system and investment flowed out. Something obviously goes on inside the financial system to channel the savings into investment, and it is time we learned just what this something is.

There is another, equally important, reason for studying the financial system. *Fiscal policy* is not the only lever the government has on the economy's aggregate demand curve; it also exercises significant control over aggregate demand by manipulating *monetary policy.* If we are to understand monetary policy (the subject of Chapters 30 and 31), we must first acquire some understanding of the financial system.

The present chapter has three major objectives. It first seeks to explain the nature of money: what it is, what purposes it serves, and how it is measured. Once this is done, we turn our attention to the banking system, explaining its historical origins, the nature of banking as a business, and why this industry is so heavily regulated. Finally, we learn how banks create money—a subject that is of great importance because you cannot hope to understand monetary policy until you first understand how money is created.

At the end of the chapter, we will see why government authorities must exercise control over the supply of money in a modern economy. This leads naturally into next chapter's discussion of *central banking,* that is, the techniques used to implement monetary policy.

ISSUE: TO REGULATE OR TO DEREGULATE? THAT IS THE QUESTION

The pendulum of bank regulation has swung back and forth over the last 2 decades. In the late 1970s and early 1980s, the United States eased several restrictions on interest rates and permissible activities for banks. Then, after a number of banks and a much larger number of savings institutions went bankrupt in the

1980s, Congress and the bank regulatory agencies cracked down with stiffer regulation and closer scrutiny. In the last few years, the pendulum has swung back toward deregulation again. Restrictions on banking across state lines were lifted in 1994 and regulators have expanded the list of activities permissible for banks.

How much bank regulation is enough? To make an informed judgment on this issue, we must first ask a more basic question: Why were banks so heavily regulated in the first place?

Banking is certainly not heavily monopolized. While there are financial giants such as Citibank (New York) and Bank of America (California), the industry is populated by literally thousands of small banks located in cities and towns throughout the country. There are over 10,000 commercial banks and several thousand savings institutions nationwide. So why did government regulations formerly tell banks, to some degree, how much they could accept in deposits, how much interest they could pay on these deposits, what types of investments they could make, and so on?

A first reason is that the major "output" of the banking industry—the nation's supply of money—is an important determinant of aggregate demand, as we will see in Chapter 30. Bank managers presumably do what is best for their stockholders. That, at any rate, is their job. But as we shall see, what is best for bank stockholders may not be best for the whole economy. For this reason, the government does not allow bankers to determine the level of the nation's money supply by profit considerations alone.

A second reason for the extensive regulation of banks is concern for the safety of depositors. In a free-enterprise system, new businesses are born and die every day; and no one other than the people immediately involved takes much notice of these goings-on. When a firm goes bankrupt, stockholders lose money and employees may lose their jobs. (The latter may not even happen if new management takes over the assets of the bankrupt firm.) But, except for the case of very large firms, that is about it.

A **RUN ON A BANK** occurs when many depositors withdraw cash from their accounts all at once.

But banking is different. If banks were treated like other firms, depositors would lose money whenever one went bankrupt. That is bad enough by itself, but the real danger comes in the case of a **run on a bank.** When depositors get jittery about the security of their money, they may all rush to cash in their accounts. For reasons we will learn in this chapter, most banks could not survive such a "run" and would be forced into insolvency.

Worse yet, this disease is highly contagious. If Mr. Smith hears that his neighbor has just lost her life savings because the Main Street National Bank went broke, he is quite likely to rush to his own bank to make a hefty withdrawal. In fact, that is precisely what happened to a number of savings banks in the 1980s.

Without modern forms of bank regulation, therefore, one bank failure might lead to another. Indeed, bank failures have been common throughout most of American history. (See Figure 29-1.) But, since the 1930s, bank failures generally have not been precipitated by runs because the government has taken steps to ensure that such an infectious disease, if it occurs, will not spread. It has done this in several ways that we will mention in this chapter.

BARTER VERSUS MONETARY EXCHANGE

Money is so much a part of our daily existence that we take it for granted and fail to appreciate all that it accomplishes. But money is in no sense "natural."

BARTER is a system of exchange in which people directly trade one good for another, without using money as an intermediate step.

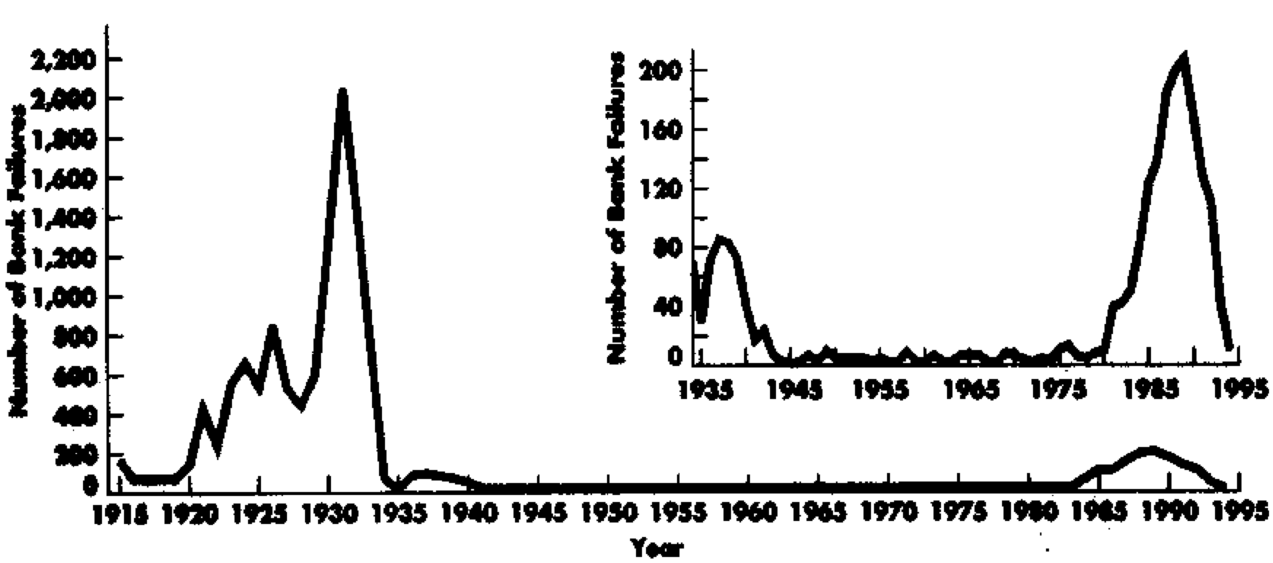
Like the wheel, it had to be invented. The most obvious way to trade commodities is not by using money, but by **barter**—a system in which people exchange one good directly for another. And the best way to appreciate what monetary exchange accomplishes is to imagine a world without it.

Under a system of direct barter, if Farmer Jones grows corn and has a craving for peanuts, he has to find a peanut farmer, say, Farmer Smith, with a taste for corn. If he finds such a person (this was called the *double coincidence of wants* by the classical economists), they make the trade. If this sounds easy, try to imagine how busy Farmer Jones would be if he had to repeat the sequence for every commodity he consumed in a week. For the most part, the desired double coincidences of wants are more likely to turn out to be double wants of coincidence. Jones gets no peanuts and Smith gets no corn. Worse yet, with so much time spent looking for trading partners, Jones would have far less time to grow corn. In brief:

Money greases the wheels of exchange, and thus makes the whole economy more productive.

Under a monetary system, Farmer Jones gives up his corn for money. He does so not because he wants the money per se, but because of what that money can buy. Money makes his shopping tasks much easier, allowing him simply to locate a peanut farmer who wants money. And what peanut farmer does not? For these reasons, monetary exchange replaced barter at a very early stage of human civilization, and only extreme circumstances, like massive wars and runaway inflations, have been able to bring barter (temporarily) back.

FIGURE 29-1 BANK FAILURES IN THE UNITED STATES, 1915–1994



This chart shows the number of commercial banks that failed each year from 1915 through 1994. Notice the sharp drop in the number of failures from 1932 to 1934 and the spurt in the 1980s. In general, however, failures clearly are much less common in the postwar period than they were in earlier years.

SOURCE: Federal Deposit Insurance Corporation.

THE CONCEPTUAL DEFINITION OF MONEY

MONEY is the standard object used in exchanging goods and services. In short, money is the **MEDIUM OF EXCHANGE.**

The **UNIT OF ACCOUNT** is the standard unit for quoting prices.

A **STORE OF VALUE** is an item used to store wealth from one point in time to another.

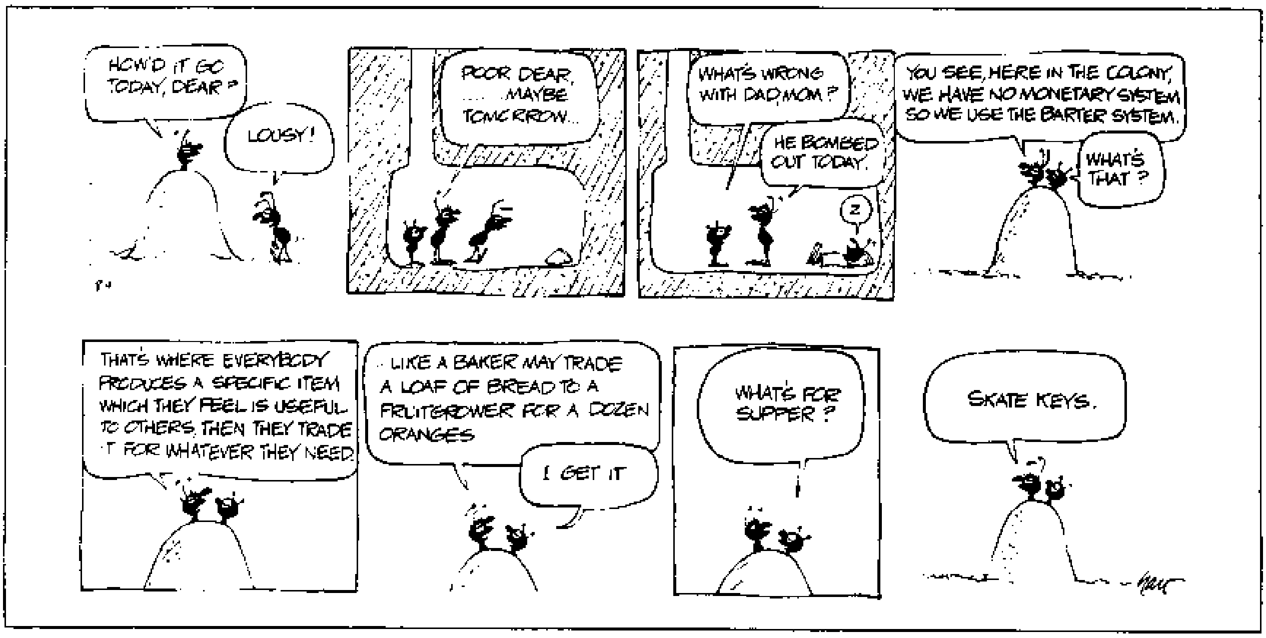
Monetary exchange is the alternative to barter. In a system of monetary exchange, people trade **money** for goods when they purchase something and trade goods for money when they sell something; but they do not trade goods directly for other goods. This defines money's principal role as the **medium of exchange.** But once it has become accepted as the medium of exchange, whatever serves as money is bound to take on other functions as well. For one, it will inevitably become the **unit of account,** that is, the standard unit for quoting prices. Thus, if inhabitants of an idyllic tropical island use coconuts as money, they would be foolish to quote prices in terms of sea shells.

Money may also come to be used as a **store of value.** If Farmer Jones' corn sales bring in more value than he wants to spend right away, he may find it convenient to store the difference temporarily in the form of money. This is because he knows that money can be "sold" easily for goods and services at a later date, whereas land, gold, and other stores of value might not be. Of course, if money pays no interest and inflation is substantial, he may decide to forgo the convenience of money and store his wealth in some other form rather than see its purchasing power eroded. So this role of money is far from inevitable.

Since money may not always serve as a store of value, and since there are stores of value other than money, it is best not to include the store-of-value function as part of our conceptual definition of money. Instead, we simply label as "money" whatever serves as the medium of exchange.

WHAT SERVES AS MONEY?

Anthropologists and historians will testify that a bewildering variety of objects have served as money in different times and places. Cattle, stones, candy bars, cigarettes, woodpecker scalps, porpoise teeth, and giraffe tails are a few of the more colorful examples.



A **COMMODITY MONEY** is an object in use as a medium of exchange, but which also has a substantial value in alternative (nonmonetary) uses.

In primitive or less organized societies, the commodities that served as money generally had value in themselves. If not used as money, cattle could be slaughtered for food, cigarettes could be smoked, and so on. But such **commodity money** generally runs into several severe difficulties. To be useful as a medium of exchange, the commodity must be divisible. This makes cattle a poor choice. It must also be of uniform, or at least readily identifiable, quality so that inferior substitutes are easy to recognize. This may be why woodpecker scalps never achieved great popularity. The medium of exchange must also be storable and durable, which presents a serious problem for candy-bar money. Finally, because commodity money needs to be carried and stored, it is helpful if the item is compact, that is, has high value per unit of volume and weight.

All of these traits make it sensible that gold and silver have circulated as money since the first coins were struck about 2,500 years ago. As they have high value in nonmonetary uses, a lot of purchasing power can be carried without too much weight. Pieces of gold are also storable, divisible (with a little trouble), and of identifiable quality (with a little more trouble).

The same characteristics suggest that paper would make an ideal money. The Chinese invented paper money in the 11th century, and Marco Polo brought the idea to Europe. Since we can print any number on it that we please, we can make paper money as divisible as we like and also make it possible to carry a large value in a lightweight and compact form. Paper is easy to store and, with a little cleverness, we can make counterfeiting hard (though never impossible).

Paper cannot, however, serve as commodity money because its value per square inch in alternative uses is so small. A paper currency that is repudiated by its issuer can, perhaps, be used as wallpaper or to wrap fish, but these uses will surely represent only a small fraction of the paper's value as money.[1] Contrary to the popular expression, such a currency literally *is* worth the paper it is printed on—which is to say that it is not worth much. Thus, paper money is always **fiat money**.

FIAT MONEY is money that is decreed as such by the government. It is of little value as a commodity, but it maintains its value as a medium of exchange because people have faith that the issuer will stand behind the pieces of printed paper and limit their production.

Money in the contemporary United States is almost entirely fiat money. Look at a dollar bill. Next to George Washington's picture it states: "This note is legal tender for all debts, public and private." Nowhere on the certificate is there a promise, stated or implied, that the U.S. government will exchange it for anything else. A dollar bill is convertible into, say, 4 quarters or 10 dimes—but not into gold, chocolate, or any other commodity.

Why do people hold these pieces of paper? Only because they know that others are willing to accept them for things of intrinsic value—food, rent, shoes, and so on. If this confidence ever evaporated, these dollar bills would cease serving as a medium of exchange and, given that they make ugly wallpaper, would become virtually worthless.

But don't panic. This is not likely to occur. Our current monetary system has evolved over hundreds of years during which *commodity* money was first replaced by *full-bodied paper money*—paper certificates that were backed by gold or silver of equal value held in the issuer's vaults. Then the full-bodied paper money was replaced by certificates that were only partially backed by gold and silver. Finally, we arrived at our present system, in which paper money has no "backing" whatsoever. Like a hesitant swimmer who first dips her toes, then her

[1]The first paper money issued by the federal government, the Continental dollar, was essentially repudiated. (Actually, the new government of the United States redeemed the Continentals for 1 cent on the dollar in the 1790s.) This gave rise to the derisive expression, "It's not worth a Continental."

legs, then her whole body into a cold swimming pool, we have "tested the water" at each step of the way—and found it to our liking. It is unlikely that we will ever take a step back in the other direction.

HOW THE QUANTITY OF MONEY IS MEASURED

Since the amount of money in circulation is of profound importance for the determination of national product and the price level, it is important that the government know how much money there is. So we must devise some *measure* of the money supply.

Our conceptual definition of money describes it as the medium of exchange. But this raises difficult questions about just what items to include and exclude when we count up the money supply—questions that have long made the statistical definition of money a subject of dispute. In fact, the U.S. government has several official definitions of the money supply, two of which we will meet shortly.

Some components are obvious. All of our coins and paper money, the small change of our economic system, clearly should count as money. But we cannot stop there if we want to include the main vehicle for making payments in our society, for the lion's share of our nation's payments are made neither in metal nor in paper money, but by check.

Checking deposits are actually no more than bookkeeping entries in bank ledgers. Many people think of checks simply as a convenient way to give coins or dollar bills to someone else. But that is not so. In fact, the volume of money held in the form of checkable deposits far exceeds the volume of currency. For example, when you pay the grocer $50 by check, dollar bills rarely change hands. Instead, that check normally travels back to your bank, where $50 is deducted from the bookkeeping entry that records your account and added to the bookkeeping entry for your grocer's account. (If you and the grocer hold accounts at different banks, more books get involved; but still no coins or bills are likely to be moved.)

Since so many transactions are made by check, it seems imperative that checkable deposits be included in any useful definition of the money supply. Unfortunately, this is not an easy task nowadays because of the wide variety of ways to transfer money by check. Traditional checking accounts in commercial banks are the most familiar. But many people can also write checks on their savings accounts, on their deposits at credit unions, on their mutual funds, their accounts with stockbrokers, and so on.

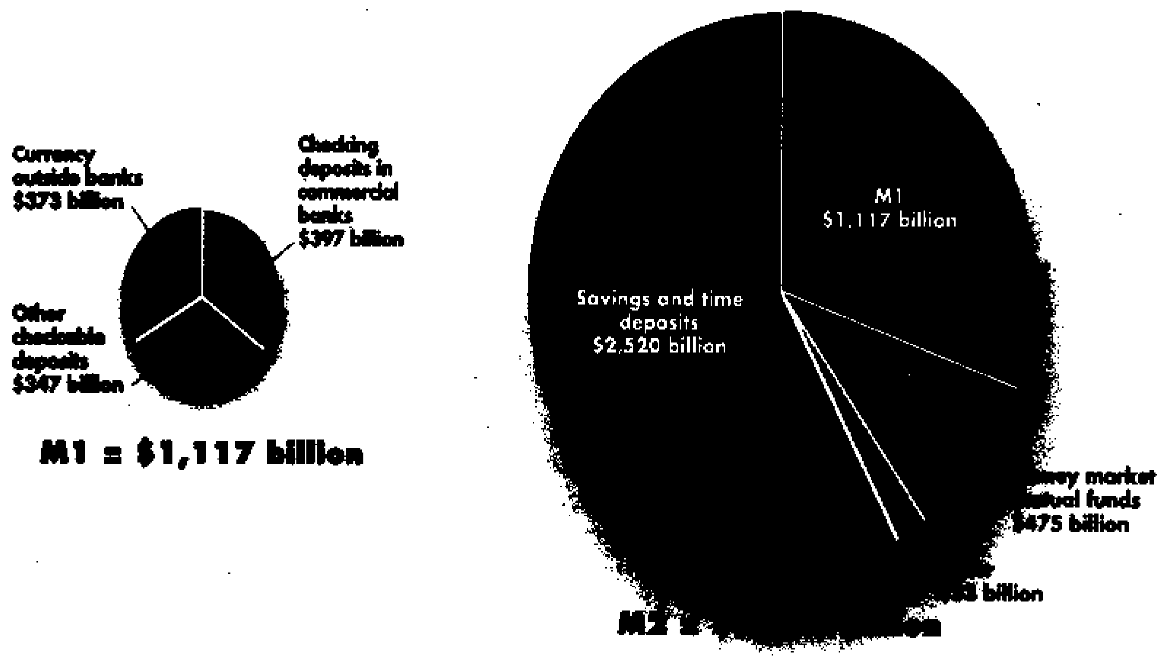
The narrowly defined money supply, usually abbreviated **M1**, is the sum of all coins and paper money in circulation, plus certain checkable deposit balances at banks and savings institutions.[2]

One popular definition of the money supply draws the line early and includes only coins, paper money, travelers' checks, conventional checking accounts, and certain other checkable deposits in banks and savings institutions. In the official U.S. statistics, this narrowly defined concept of money is called **M1**. The left-hand side of Figure 29-2 shows the composition of M1 as of February 1996.

But there are other types of accounts that allow withdrawals by check and which therefore are candidates for inclusion in the money supply. Most notably, *money market deposit accounts* allow only a few checks per month but pay market-determined interest rates. Consumers have found these accounts extremely

[2]This includes travelers' checks and NOW (negotiable order of withdrawal) accounts.

FIGURE 29-2 TWO DEFINITIONS OF THE MONEY SUPPLY (FEBRUARY 1996)



al Reserve.

attractive vehicles for short-term investment, and balances in them now exceed all the checkable deposits included in M1.

In addition, many mutual fund organizations and brokerage houses offer *money market mutual funds*. These funds sell shares and use the proceeds to purchase a variety of short-term securities. But the important point, for our purposes, is that owners of shares in money market mutual funds can make withdrawals by writing a check. So depositors can—and sometimes do—use their holdings of fund shares just like checking accounts.

Finally, although you cannot write a check on a *savings account*, most economists feel that modern banking procedures have blurred the distinction between checking balances and savings balances. For example, most banks these days offer convenient electronic transfers of funds from one account to another, either by telephone or by pushing a button on an automated teller. Consequently, savings balances can become checkable almost instantly. For this reason, savings accounts are included—along with money market deposit accounts, money market mutual fund shares, and a few other small items—in the broader definition of the money supply known as **M2**.

The broadly defined money supply, usually abbreviated **M2**, is the sum of all coins and paper money in circulation, plus all types of checking account balances, plus most forms of savings account balances, plus shares in money market mutual funds, and a few other minor items.

The composition of M2 as of February 1996 is shown on the right-hand side of Figure 29-2. You can see that savings deposits predominate, dwarfing everything that is included in M1. Figure 29-2 illustrates two points that are worth remembering. First, our money supply comes not only from banks, but also from savings institutions, brokerage houses, and mutual fund organizations. Second, however, banks still play a predominant role.

Some economists do not want to stop counting at M2; they prefer still broader definitions of money (M3, and so on) which include more types of bank deposits and other closely related assets. The inescapable problem, however, is that there

NEAR MONEYS are liquid assets that are close substitutes for money.

An asset's **LIQUIDITY** refers to the ease with which it can be converted into cash.

is no obvious place to stop, no clear line of demarcation between those assets that *are* money and those that are merely *close substitutes* for money—so-called **near moneys.**

If we define an asset's **liquidity** as the ease with which it can be converted into cash, there is a range of assets of varying degrees of liquidity. Everything in M1 is completely "liquid," the money market fund shares and passbook savings accounts included in M2 are a bit less so, and so on, until we encounter such things as short-term government bonds, which, while still quite liquid, would not normally be included in the money supply. Any number of different Ms can be defined—and have been—by drawing the line in different places.

And there are still more complexities. For example, credit cards clearly serve as a medium of exchange. So should they be included in the money supply? Yes, you say. But how much money does your credit card represent? Is it the amount you currently owe on the card, which may well be zero? Or is it your entire line of credit, even though you may never use it all? Neither seems a sensible choice, which is one reason why economists have so far ignored credit cards in the definition of money. And soon Americans may start using electronic money instead of cash. Money will be transferred via computer hookups or by so-called *smart cards* with memory chips. (See the box on the next page.)

There are further definitional issues that we have not mentioned. But an introductory course in economics is not the place to get bogged down in complex definitional issues. So we will simply adhere to the convention that *"money" consists only of coins, paper money, and checkable deposits.*

Now that we have defined money and seen how it can be measured, we turn our attention to the principal creators of money—the banks.

HOW BANKING BEGAN

When Adam and Eve left the Garden of Eden, they did not encounter a branch of Citibank. Banking had to be invented, and some time passed before it came to be practiced as it is today. With a little imagination, we can see how the first banks must have begun.

When money was made of gold, it was most inconvenient for consumers and merchants to carry it around and weigh and assay it for purity every time a transaction was made. So the practice developed of leaving one's gold in the safe storage facilities of a goldsmith and carrying in its place a receipt stating that John Doe did indeed own 5 ounces of gold. When people began trading goods and services for the goldsmiths' receipts, rather than for the gold itself, the receipts became an early form of paper money.

At this stage, paper money was fully backed by gold. But gradually the goldsmiths began to notice that the amount of gold they were actually required to pay out in a day was but a small fraction of the total gold they had stored in their warehouses. Then one day some enterprising goldsmith hit upon a momentous idea that must have made him fabulously wealthy.

His thinking probably ran something like this. "I have 2,000 ounces of gold stored away in my vault, for which I collect storage fees from my customers. But I am never called upon to pay out more than 100 ounces on a single day. What harm could it do if I lent out, say, half the gold I now have? I'll still have more than enough to pay off any depositors that come in for a withdrawal, so no one will ever know the difference. And I could earn 30 additional ounces of gold each year in interest on the loans I make (at 3 percent interest on 1,000 ounces).

IS THERE A SMART CARD IN YOUR FUTURE?



In the mid-1990s, several banks and other companies began test marketing new forms of high-tech electronic currency or "e-cash." One proposed form uses encrypted electronic messages to send balances from one computer to another, possibly bypassing banks entirely.

But the most commonly discussed form of e-cash, at least so far, is the so-called *stored-value card*. It works as follows. Soon you may be able to buy a "smart card" (which looks much like a credit card) with an embedded memory chip on which your initial payment is recorded. At this point, your conventional money is transformed into electronic currency. Thereafter, you can use the cash stored on the card to make purchases simply by inserting your card into specially designed slots in vending machines and stores. When the stored value on your card gets depleted, you can replenish it at an ATM.

Electronic currency raises several novel issues. For one thing, the federal government has long held a monopoly over currency issue; e-cash may erode that monopoly. For another, consumers and businesses may have privacy and safety concerns: Will their electronic transactions be safe from system errors and computer snoopers and hackers? Law enforcement agencies are also worried that large, untraceable electronic transfers of funds may be a draw to criminals and tax evaders.

While futurists confidently predict that these new products represent the money of the future, skeptics note that we have heard such claims before—and they have never come true. Only time will tell if the new technologies will catch on.

With this profit, I could lower my service charges to depositors and so attract still more deposits. I think I'll do it."

FRACTIONAL RESERVE BANKING is a system under which bankers keep as reserves only a fraction of the funds they hold on deposit.

With this resolution, the modern system of **fractional reserve banking** was born. This system has three features that are crucially important to this chapter.

1. *Bank profitability.* By getting deposits at zero interest and lending some of them out at positive interest rates, goldsmiths made a profit. The history of banking as a profit-making industry was begun and has continued to this date. *Banks, like other enterprises, are in business to earn profits.*
2. *Bank discretion over the money supply.* When goldsmiths decided to keep only fractions of their total deposits on reserve and lend out the balance, they acquired the ability to *create money*. As long as they kept 100 percent reserves, each gold certificate represented exactly 1 ounce of gold. So whether people decided to carry their gold or leave it with their goldsmith did not affect the money supply, which was set by the volume of gold.

   With the advent of fractional reserve banking, however, new paper certificates were added whenever goldsmiths lent out some of the gold they held on deposit. The loans, in effect, created new money. In this way, the total amount of money came to depend on the amount of gold that each goldsmith felt compelled to keep in his vault. For any given volume of gold on deposit, the lower the reserves the goldsmiths kept, the more loans they could make, and therefore the more money there would be. While we no longer use gold to back our money, this principle remains true today. *Bankers' business decisions influence the supply of money.*
3. *Exposure to runs.* A goldsmith who kept 100 percent reserves never had to worry about a run on his vault. Even if all his depositors showed up at the door at once, he could always convert their paper receipts back into gold. But as soon as the first goldsmith decided to get by with only fractional reserves, the possibility of a run on the vault became a real concern. If that first goldsmith who lent out half his gold had found 51 percent of his customers at his door one unlucky day, he would have had a lot of explaining

to do. Similar problems have worried bankers for centuries. *The danger of runs on the bank has induced bankers to keep prudent reserves and to lend out money carefully.*

PRINCIPLES OF BANK MANAGEMENT: PROFITS VERSUS SAFETY

Bankers have a reputation for conservatism in politics, dress, and business affairs. From what has been said so far, the economic rationale for this conservatism should be clear. Checking deposits are pure fiat money. Years ago, these deposits were "backed" by nothing more than the bank's promise to convert them into currency on demand. If people lost trust in a bank, the bank was doomed.

Thus, it has always been imperative for bankers to acquire a reputation for prudence. This they did in two principal ways. First, they had to maintain a sufficiently generous level of reserves to minimize their vulnerability to runs. Second, they had to be somewhat cautious in making loans and investments, since any large losses on their loans would undermine the confidence of depositors.

It is important to realize that banking under a system of fractional reserves is an inherently risky business that is rendered safe only by cautious and prudent management. America's continuing history of bank failures bears sober testimony to the fact that many bankers have been neither cautious nor prudent. Why? Because this is not a recipe for high profits. Bank profits are maximized by keeping reserves as low as possible, by making at least some risky investments, and by granting loans to borrowers of questionable credit standing who will pay high interest rates.

The art of bank management is to strike the appropriate balance between the lure of profits and the need for safety. If a banker errs by being too stodgy, his bank will earn inadequate profits. If he errs by taking unwarranted risks, his bank may not survive at all. Many banks have perished in the latter way in recent years, especially in the savings and loan industry. (See the box on the next page.)

BANK REGULATION

Governments in virtually every society have decided that the balance between profits and safety likely to be preferred by profit-minded bankers may not be where society wants it struck. So they have thrown up a web of regulations designed to insure the safety of depositors and to control the supply of money.

DEPOSIT INSURANCE is a system that guarantees that depositors will not lose money even if their bank goes bankrupt.

The principal innovation guaranteeing the safety of bank deposits is **deposit insurance.** Today most bank deposits are insured against loss by the *Federal Deposit Insurance Corporation (FDIC)*—an agency of the U.S. government. If your bank belongs to the FDIC (and almost all do), your account is insured for up to $100,000 regardless of what happens to the bank. Thus, while bank failures may spell disaster for the bank's stockholders, they do not give many depositors cause for concern. Deposit insurance eliminates the motive for customers to rush to their bank just because they hear some bad news about the bank's finances. Many observers give this innovation much of the credit for the pronounced decline in bank failures after 1933, the year the FDIC was established. (Refer back to Figure 29-1 on page 673.)

In addition to insuring depositors against loss, the government takes steps to see that banks do not get into financial trouble. For one thing, various regulatory authorities conduct periodic *bank examinations* in order to keep tabs on the

THE SAVINGS AND LOAN CRISIS: OVER AT LAST

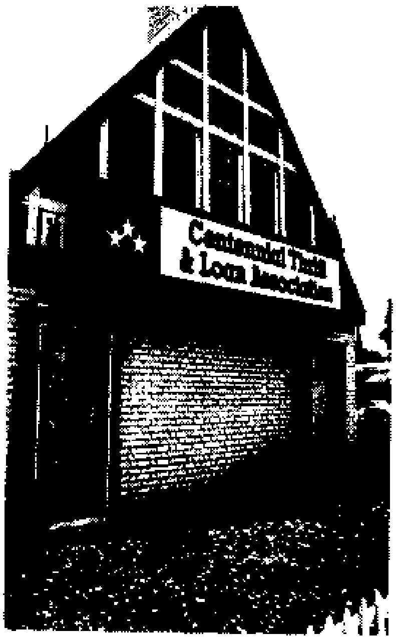
At last, it's over! The Resolution Trust Corporation (RTC), the agency Congress created to clean up the debris left by the collapse of America's savings and loan industry, closed its doors at the end of 1995, having merged or sold off the assets of approximately 750 thrift institutions. The final bill to the American taxpayer has been estimated at about $150 billion.

How did this mess happen? Like many debacles, it took both bad luck and bad policy. The bad luck started in the 1970s and early 1980s, when interest rates skyrocketed. S&Ls found themselves stuck with old mortgages carrying low interest rates while paying high current interest rates to depositors—a losing proposition.

The bad policy came in two waves. First came the financial deregulation of the 1980s, which eased supervision just as thrifts were allowed to expand their lending beyond their traditional domain of home mortgages. The result was ill-advised investments, imprudent risk-taking, and an outrageous amount of fraud. With their institutions teetering on the brink, many S&L executives decided to "bet the bank" on risky ventures that, they hoped, would pay off. After all, if the investments went sour, depositors would be protected by deposit insurance.

Then came the second mistake: procrastination in Washington. The broad outlines of the S&L problem were apparent by the mid-1980s, but Congress and the Reagan administration temporized rather than face up to the costs of closing the moribund thrifts. Politicians feared (correctly?) that taxpayers would be angry when they received the huge bill for deposit insurance. Only in 1989, when the problem had reached crisis dimensions, did President Bush persuade Congress to bite the bullet and create the RTC to finish the job.

The lesson of the thrift debacle was a painful one, not likely to be forgotten soon. It led, among other things, to tighter regulations, stiffer penalties for fraud, and higher premiums for deposit insurance (which were just reduced in 1995)—all designed to make sure it never happens again.



financial conditions and business practices of the banks under their purview. Bank supervision was tightened in 1992 by legislation which permits the authorities to intervene early in the affairs of financially troubled banks. There are also laws and regulations that *limit the kinds and quantities of assets in which banks may invest.* For example, banks are permitted to own only limited amounts of common stock. Both of these forms of regulation are clearly aimed at maintaining bank safety.

REQUIRED RESERVES are the minimum amount of reserves (in cash or the equivalent) required by law. Normally, required reserves are proportional to the volume of deposits.

A final type of regulation also has some bearing on safety but is motivated primarily by the government's desire to control the money supply. We have seen that the amount of money any bank will issue depends on the amount of reserves it elects to keep. For this reason, most banks are subject by law to minimum **required reserves.** While banks may (and sometimes do) keep reserves in excess of these legal minimums, they may not keep less. It is this regulation that places an upper limit on the money supply. The rest of this chapter is concerned with the details of this mechanism.

HOW BANKERS KEEP BOOKS

Before we can fully understand the mechanics of modern banking and the process by which money is "created," we must acquire at least a nodding

POLICY DEBATE ABOLISH THE COMMUNITY REINVESTMENT ACT?

Not all bank regulation is aimed at the safety of depositors. Some regulation is designed to protect borrowers from discriminatory practices by banks. One particularly controversial law is the Community Reinvestment Act (CRA), which was designed to combat "redlining"—the practice of delineating certain geographical areas, often minority neighborhoods, in which banks did not lend.

Bankers complain that CRA regulations bury them in paperwork and force them to make unprofitable loans. Such lending amounts to government-enforced allocation of credit, critics contend. Community advocacy groups counter that access to credit, which is inadequate in low-income and moderate-income areas, is crucial for economic success. A strong CRA keeps the pressure on banks to do better, they argue.

Because the statute is terse and somewhat vague, much depends on how bank regulators interpret it. So in 1993 President Clinton ordered the regulatory agencies to rewrite the CRA rules in ways that would focus more on *results* and less on *paperwork*. The lengthy and contentious rule-making process took more than 2 years. Then, after the 1994 elections, Republicans in Congress introduced several legislative proposals that would essentially repeal the CRA, thereby relieving banks of the regulatory burden and letting the market work. President Clinton threatened to veto any such legislation, and the idea died in Congress in 1995.

But the controversy continues. Should the government pursue what some people consider an affirmative action policy for bank lending? Or should such matters be left to the market? What do you think?

acquaintance with the way bankers keep their books. The first thing to know is how to distinguish assets from liabilities.

An **ASSET** of an individual or business firm is an item of value that the individual or firm owns.

A **LIABILITY** of an individual or business firm is an item of value that the individual or firm owes. Many liabilities are known as *debts*.

An **asset** of a bank is something of value that the bank *owns*. This "thing" may be a physical object, such as the bank building or a computer, or it may be just a piece of paper, such as an IOU of a customer to whom the bank has made a loan. A **liability** of a bank is something of value that the bank *owes*. Most bank liabilities take the form of bookkeeping entries. For example, if you have an account in the Main Street Bank, your bank balance is a liability of the bank. (It is, of course, an asset to you.)

There is an easy test to see whether some piece of paper or bookkeeping entry is a bank's *asset* or *liability*. Ask yourself whether, if this paper were converted into cash, the bank would receive the cash (if so, it is an asset) or pay it out (if so, it is a liability). This test makes it clear that loans to customers are bank assets (when loans are repaid, the bank collects), while customers deposits are bank liabilities (when deposits are cashed in, the bank must pay). Of course, things are just the opposite to the bank's customers: The loans are liabilities and the deposits are assets.

A **BALANCE SHEET** is an accounting statement listing the values of all the assets on the left-hand side and the values of all the liabilities *and net worth* on the right-hand side.

NET WORTH is the value of all assets minus the value of all liabilities.

When accountants draw up a complete list of all the bank's assets and liabilities, the resulting document is called the bank's **balance sheet.** Typically, the value of all the bank's assets exceeds the value of all its liabilities. (On the rare occasions when this is not so, the bank is in serious trouble.) In what sense, then, do balance sheets "balance"?

They balance because accountants have invented the concept of **net worth** to balance the books. Specifically, they define the net worth of a bank to be the difference between the value of all its assets and the value of all its liabilities. Thus, by definition, when accountants add net worth to liabilities, the sum they get must be the same as the value of the bank's assets. In short:

$$\text{Assets} = \text{Liabilities} + \text{Net Worth}$$

Table 29-1 illustrates this with the balance sheet of a fictitious bank, Bank-a-mythica, whose finances are extremely simple. On December 31, 1996, it had only

TABLE 29-1 BALANCE SHEET OF BANK-A-MYTHICA, DECEMBER 31, 1996

Assets		Liabilities and Net Worth	
Assets		**Liabilities**	
Reserves	$1,000,000	Checking deposits	$5,000,000
Loans outstanding	4,500,000		
Total	$5,500,000	**Net Worth**	
Addendum: Bank Reserves		Stockholders' equity	500,000
Actual reserves	$1,000,000	Total	$5,500,000
Required reserves	1,000,000		
Excess reserves	0		

two kinds of assets (listed on the left-hand side of the balance sheet)—$1.0 million in cash, which it held as reserves, and $4.5 million in outstanding loans to its customers, that is, in customers' IOUs. And it had only one type of liability (listed on the right-hand side)—$5.0 million in checking deposits. The difference between total assets ($5.5 million) and total liabilities ($5.0 million) was the bank's net worth ($500,000), shown on the right-hand side of the balance sheet.

THE LIMITS TO MONEY CREATION BY A SINGLE BANK

Let us now turn to the process of deposit creation. Many bankers will deny that they have any ability to "create" money. The phrase itself has a suspiciously hocus-pocus sound to it. But they are not quite right. For although any individual bank's ability to create money is severely limited, the banking system as a whole can achieve much more than the sum of its parts. Through the modern alchemy of *deposit creation*, it can turn one dollar into many dollars. But to understand this important process, we had better proceed in steps, beginning with the case of a single bank, our hypothetical Bank-a-mythica.

According to the balance sheet in Table 29-1, Bank-a-mythica is holding cash reserves of $1 million, equal to 20 percent of its $5 million in deposits. Assume that this is the minimum reserve ratio prescribed by law and that the bank strives to keep its reserves down to the legal minimum; that is, it strives to keep its **excess reserves** down to zero.

EXCESS RESERVES are any reserves held in excess of the legal minimum.

Now let us suppose that on January 2, 1997, an eccentric widower comes into Bank-a-mythica and deposits $100,000 in cash in his checking account. The bank now has $100,000 more in cash reserves, and $100,000 more in checking deposits. But since deposits are up by $100,000, *required* reserves are up by only $20,000, leaving $80,000 in *excess* reserves. Table 29-2 illustrates the effects of this transaction on Bank-a-mythica's balance sheet. Tables such as this, which show *changes* in balance sheets rather than the balance sheets themselves, will help us follow the money-creation process.[3]

[3]Notice that in all such tables, which are called *T accounts*, the two sides of the ledger must balance. This is because changes in assets and changes in liabilities must be equal if the balance sheet is to balance both before and after the transaction.

TABLE 29-2 CHANGES IN BANK-A-MYTHICA'S BALANCE SHEET, JANUARY 2, 1997

Assets		Liabilities	
Reserves	+$100,000	Checking deposits	+$100,000
Addendum: Changes in Reserves			
Actual reserves	+$100,000		
Required reserves	+ $20,000		
Excess reserves	+ $80,000		

Bank-a-mythica receives a $100,000 cash deposit. It now holds excess reserves of $80,000, since required reserves rise by only $20,000 (20 percent of $100,000).

Bank-a-mythica is unlikely to be happy with the situation illustrated in Table 29-2, for it is holding $80,000 in excess reserves on which it earns no interest. So as soon as possible it will lend out the extra $80,000—let us say to Hard-Pressed Construction Company. This loan leads to the balance sheet changes shown in Table 29-3: Bank-a-mythica's loans rise by $80,000 while its holdings of cash reserves fall by $80,000.

By combining Tables 29-2 and 29-3, we arrive at Table 29-4, which summarizes all the bank's transactions for the week. Reserves are up $20,000, loans are up $80,000, and, now that the bank has had a chance to adjust to the inflow of deposits, it no longer holds excess reserves.

Looking at Table 29-4 and keeping in mind our specific definition of money, it appears at first that the chairman of Bank-a-mythica is right when he claims not to have engaged in the nefarious practice of "money creation." All that happened was that, in exchange for the $100,000 in cash it received, the bank issued the widower a checking balance of $100,000. This does not change M1; it merely converts one form of money into another.

But wait. What happened to the $100,000 in cash that the eccentric man brought to the bank? The table shows that $20,000 was retained by Bank-a-mythica in its vault. Since this currency is no longer in circulation, it no longer counts in the official money supply. (Notice that Figure 29-2 included only "currency

TABLE 29-3 CHANGES IN BANK-A-MYTHICA'S BALANCE SHEET, JANUARY 3–6, 1997

Assets		Liabilities
Loans outstanding	+$80,000	No change
Reserves	−$80,000	
Addendum: Changes in Reserves		
Actual reserves	−$80,000	
Required reserves	No change	
Excess reserves	−$80,000	

Bank-a-mythica gets rid of its excess reserves by making a loan of $80,000 to Hard-Pressed Construction Company.

TABLE 29-4 CHANGES IN BANK-A-MYTHICA'S BALANCE SHEET, JANUARY 2–6, 1997

Assets		Liabilities	
Reserves	+$20,000	Checking deposits	+$100,000
Loans outstanding	+$80,000		
Addendum: Changes in Reserves			
Actual reserves	+$20,000		
Required reserves	+$20,000		
Excess reserves	No change		

When it receives $100,000 in cash deposits, Bank-a-mythica keeps only the required $20,000 in reserves and lends out the remaining $80,000 to Hard-Pressed Construction Company. Its excess reserves return to zero.

outside banks.") But the other $80,000, which the bank lent out, is still in circulation. It is held by Hard-Pressed Construction, which probably will redeposit it in some other bank. But even before this happens, the original $100,000 in cash has supported a rise in the money supply: there is now $100,000 in checking deposits and $80,000 in cash in circulation, making a total of $180,000. The money creation process has begun.

MULTIPLE MONEY CREATION BY A SERIES OF BANKS

Let us now trace the $80,000 in cash and see how the process of money creation gathers momentum. Suppose that Hard-Pressed Construction Company, which banks across town at the First National Bank, deposits the $80,000 into its bank account. First National's reserves increase by $80,000. But because deposits are up by $80,000, *required* reserves rise by only 20 percent of this amount or $16,000. If the management of First National Bank behaves like that of Bank-a-mythica, the $64,000 of excess reserves will be lent out.

Table 29-5 shows the effects of these events on First National Bank's balance sheet. (The preliminary steps corresponding to Tables 29-2 and 29-3 are not

TABLE 29-5 CHANGES IN FIRST NATIONAL BANK'S BALANCE SHEET

Assets		Liabilities	
Reserves	-$16,000	Checking deposits	+$80,000
Loans outstanding	+$64,000		
Addendum: Changes in Reserves			
Actual reserves	-$16,000		
Required reserves	+$16,000		
Excess reserves	No change		

Hard-Pressed deposits its $80,000 in First National Bank, which sets aside the required $16,000 in reserves (20 percent of $80,000) and lends $64,000 to Al's Auto Shop.

shown separately.) At this stage in the chain, the original $100,000 in cash has led to $180,000 in deposits—$100,000 at Bank-a-mythica and $80,000 at First National Bank—and $64,000 in cash, which is still in circulation (in the hands of the recipient of First National's loan—Al's Auto Shop). Thus, from the original $100,000, a total of $244,000 has been added to the money supply ($180,000 in checking deposits plus $64,000 in cash).

But, to coin a phrase, the bucks do not stop here. Al's Auto Shop will presumably deposit the proceeds from its loan into its own account at Second National Bank, leading eventually to the balance sheet adjustments shown in Table 29-6 when Second National makes an additional loan rather than hold on to excess reserves. You can see how the money creation process continues.

Figure 29-3 is a graphical summary of the balance sheet changes of the first five banks in the chain (from Bank-a-mythica through the Fourth National Bank), on the assumptions that each bank holds exactly the 20 percent required reserves and that each loan recipient redeposits the proceeds in the next bank. But the chain does not end there. The Main Street Movie Theatre, which received a $32,768 loan from the Fourth National Bank, then deposits these funds into the Fifth National Bank. Fifth National has to keep only 20 percent of this deposit, or $6,553.60, on reserve and will lend out the balance. And so the chain continues.

Where does it all end? The running sums in Figure 29-3 show what eventually happens to the entire banking system. The initial deposit of $100,000 in cash is ultimately absorbed in bank reserves (Column 1), leading to a total of $500,000 in new deposits (Column 2) and $400,000 in new loans (Column 3). The money supply rises by $400,000 because the nonbank public eventually holds $100,000 *less* in currency and $500,000 *more* in checking deposits.

So there really is some hocus-pocus. Somehow, an initial deposit of $100,000 leads to $500,000 in new bank deposits—a multiple expansion of $5 for every original dollar—and a net increase of $400,000 in the money supply. We had better understand why this is so. But first let us verify that the calculations in Figure 29-3 are correct.

If you look carefully at the numbers, you will see that each column forms a *geometric progression*; specifically, each entry is equal to exactly 80 percent of the entry before it. Recall that in our discussion of the multiplier in Chapter 26, we learned how to sum an infinite geometric progression, which is just what each of these chains eventually•will be. In particular, if the common ratio is R, the sum of an infinite geometric progression is:

TABLE 29-6 CHANGES IN SECOND NATIONAL BANK'S BALANCE SHEET

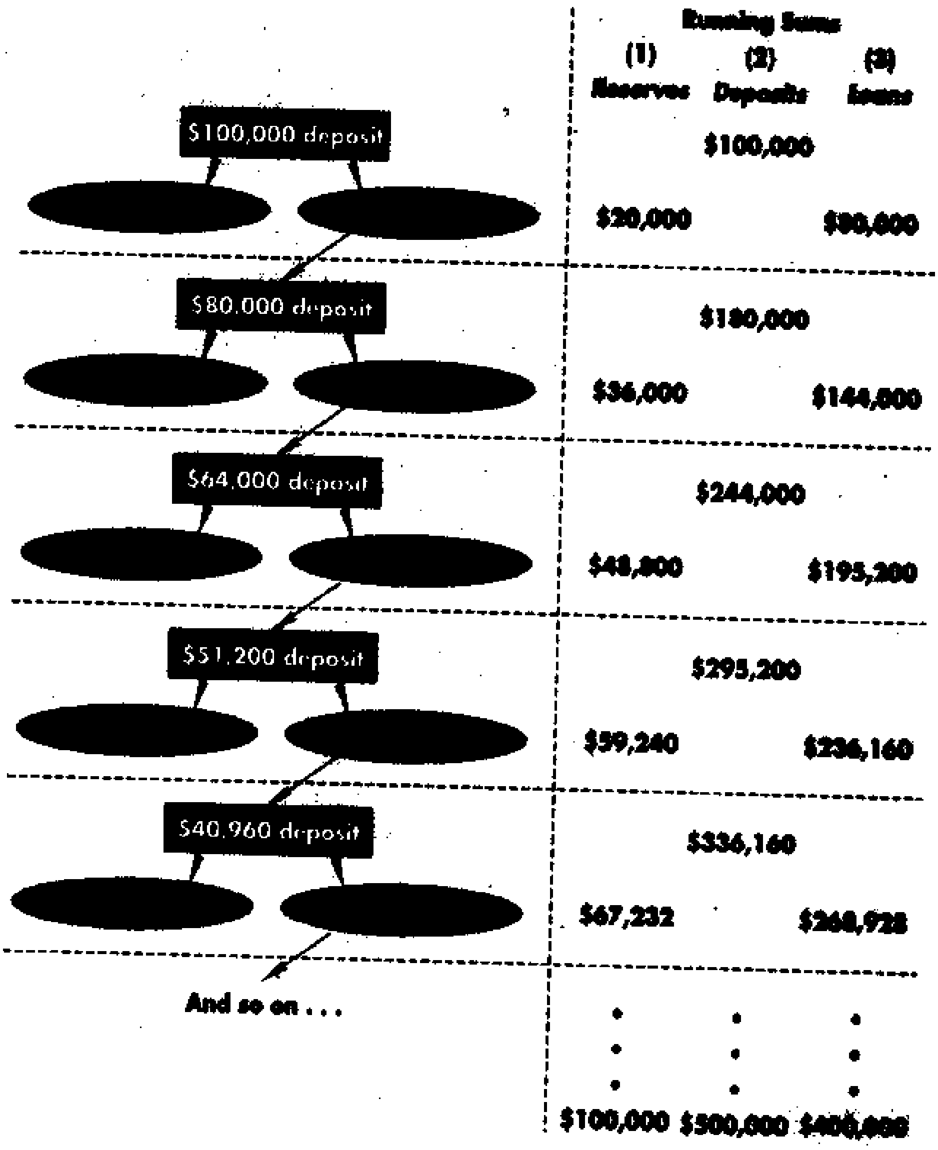
Assets		Liabilities	
Reserves	+$12,800	Checking deposits	+$64,000
Loans outstanding	+$51,200		
Addendum: Changes in Reserves			
Actual reserves	+$12,800		
Required reserves	+$12,800		
Excess reserves	No change		

When Al deposits his $64,000 in Second National Bank, that bank retains $12,800 as required reserves (20 percent of $64,000) and lends out the remaining $51,200.

FIGURE 29-3 THE CHAIN OF MULTIPLE DEPOSIT CREATION

This diagram indicates how the deposit creation process spreads through the banking system. Each bank that receives a deposit keeps 20 percent of the deposit as reserves and lends out the other 80 percent—which becomes a deposit to the next bank in the chain. And so the chain continues. The numbers on the right-hand side of the diagram indicate the running sums of reserves, deposits, and loans caused by the initial $100,000 deposit.

SOURCE: This schematic diagram was suggested to us by Dr. Ivan K. Cohen, whom we thank.



$$1 + R + R^2 + R^3 + \ldots = \frac{1}{1 - R}$$

By applying this formula to the chain of checking deposits in Figure 29-3, we get:

$$\begin{aligned} &\$100{,}000 + \$80{,}000 + \$64{,}000 + \$51{,}200 + \ldots \\ &= \$100{,}000 \times (1 + 0.80 + 0.64 + 0.512 + \ldots) \\ &= \$100{,}000 \times (1 + 0.80 + 0.80^2 + 0.80^3 + \ldots) \\ &= \$100{,}000 \times \frac{1}{1 - 0.80} = \frac{\$100{,}000}{0.20} = \$500{,}000 \end{aligned}$$

Proceeding similarly, we can verify that the new loans sum to $400,000 and that the new required reserves sum to $100,000. (Check these as exercises.) So the numbers in Figure 29-3 are correct. Let us, therefore, think through the logic behind them.

The chain of deposit creation can end only when there are no more *excess* reserves to be loaned out; that is, when the entire $100,000 in cash is tied up in *required* reserves. That explains why the last entry in Column 1 must be $100,000. But, with a reserve ratio of 20 percent, excess reserves disappear only when checking deposits expand by $500,000—which is the last entry in Column 2. Finally, since balance sheets must balance, the sum of all newly created assets (reserves plus loans) must equal the sum of all newly created liabilities ($500,000 in deposits). That leaves $400,000 for new loans—which is the last entry in Column 3.

More generally, if the reserve ratio is some number m (rather than the 1/5 in our example), each dollar of deposits requires only a fraction m of a dollar in reserves. So R, the common ratio in the above formula, is $1 - m$, and deposits must expand by $1/m$ for each dollar of new reserves that are injected into the system. This suggests the general formula for multiple deposit creation when the required reserve ratio is some number m.

OVERSIMPLIFIED DEPOSIT MULTIPLIER FORMULA

If the required reserve ratio is some fraction, m, each $1 of reserves injected into the banking system can lead to the creation of $$1/m$ in new deposits. That is, the so-called "deposit multiplier" is given by:

$$\text{Change in deposits} = \left(\tfrac{1}{m}\right) \times \text{Change in reserves}$$

Notice that this formula correctly describes what happens in our example. The initial deposit of $100,000 in cash at Bank-a-mythica constitutes $100,000 in new reserves (Table 29-2). Applying a multiplier of $1/m = 1/0.20 = 5$ to this $100,000, we conclude that bank deposits will rise by $500,000—which is just what happens. Remember, however, that the expansion process started when some eccentric widower took $100,000 in cash and deposited it in his bank account. So the public's holdings of *money*—which includes both checking deposits *and cash*—increase by only $400,000 in this case: There is $500,000 *more* in deposits, but $100,000 *less* in cash.

THE PROCESS IN REVERSE: MULTIPLE CONTRACTIONS OF THE MONEY SUPPLY

Let us now briefly consider how this deposit creation mechanism operates in reverse—as a system of deposit *destruction*. In particular, suppose that our eccentric widower came back to Bank-a-mythica to withdraw $100,000 from his checking account and return it to his mattress, where it rightfully belongs. Bank-a-mythica's *required* reserves would fall by $20,000 as a result of this transaction (20 percent of $100,000), but its *actual* reserves would fall by $100,000. The bank would be $80,000 short, as indicated in Table 29-7(a).

How does it react to this discrepancy? As some of its outstanding loans are routinely paid off, the bank will cease granting new ones until it has accumulated the necessary $80,000 in required reserves. The data for Bank-a-mythica's contraction are shown in Table 29-7(b), assuming that borrowers pay off their loans in cash.[4]

---

[4]In reality, they would probably pay with checks drawn on other banks. Bank-a-mythica would then cash these checks to acquire the reserves.

TABLE 29-7 CHANGES IN THE BALANCE SHEET OF BANK-A-MYTHIÇA

(a) Assets		(a) Liabilities		(b) Assets		(b) Liabilities
Reserves	−$100,000	Checking deposits	−$100,000	Reserves	+$80,000	No change
				Loans outstanding	−$80,000	
Addendum: Changes in Reserves				**Addendum: Changes in Reserves**		
Actual reserves	−$100,000			Actual reserves	+$80,000	
Required reserves	−$20,000			Required reserves	No change	
Excess reserves	−$80,000			Excess reserves	+$80,000	

When Bank-a-mythica loses a $100,000 deposit, it must reduce its loans by $80,000 to replenish its reserves.

But where did the borrowers get this money? Probably by making withdrawals from other banks. In this case, let us assume it all came from First National Bank, which loses an $80,000 deposit and $80,000 in reserves. It finds itself short some $64,000 in reserves [see Table 29-8(a)] and therefore must reduce its loan commitments by $64,000 [see Table 29-8(b)]. This, of course, causes some other bank to suffer a loss of reserves and deposits of $64,000, and the whole process repeats just as it did in the case of deposit expansion.

After the entire banking system had become involved, the picture would be just as shown in Figure 29-3, except that all the numbers would have *minus* signs in front of them. Deposits would shrink by $500,000, loans would fall by $400,000, bank reserves would be reduced by $100,000, and the money supply would fall by $400,000. As suggested by our deposit multiplier formula with $m = 0.20$, the decline in the bank deposits is $1/0.20 = 5$ times as large as the decline in excess reserves.

TABLE 29-8 CHANGES IN THE BALANCE SHEET OF THE FIRST NATIONAL BANK

(a) Assets		(a) Liabilities		(b) Assets		(b) Liabilities
Reserves	−$80,000	Checking deposits	−$80,000	Reserves	+$64,000	No change
				Loans outstanding	−$64,000	
Addendum: Changes in Reserves				**Addendum: Changes in Reserves**		
Actual reserves	−$80,000			Actual reserves	+$64,000	
Required reserves	−$16,000			Required reserves	No change	
Excess reserves	−$64,000			Excess reserves	+$64,000	

First National Bank's loss of the $80,000 deposit forces it to cut back its loans by $64,000.

One of the authors of this book was a student in Cambridge, Massachusetts, during the height of the radical student movement of the late 1960s. One day a circular appeared urging citizens to withdraw all funds from their checking accounts on a prescribed date, hold them in cash for a week, and then redeposit them. This act, the circular argued, would surely wreak havoc upon the capitalist system. Obviously, some of these radicals were well-schooled in modern money mechanics, for the argument was basically correct. The tremendous multiple contraction of the banking system and consequent multiple expansion that a successful campaign of this sort could have caused might have seriously disrupted the local financial system. But history records that the appeal met with little success. Checking-account withdrawals are not the stuff of which revolutions are made.

WHY THE DEPOSIT CREATION FORMULA IS OVERSIMPLIFIED

So far, our discussion of the process of money creation has made it all seem rather mechanical. If all proceeds according to formula, each $1 in new excess reserves will lead to a $\$1/m$ increase in new deposits. But in reality things are not this simple. Just as in the case of the expenditure multiplier, the oversimplified formula for deposit creation is accurate only under very particular circumstances. These circumstances require that:

1. Every recipient of cash must redeposit the cash into another bank rather than hold it.
2. Every bank must hold reserves no larger than the legal minimum.

The "chain" diagram in Figure 29-3 shows clearly what happens if either of these assumptions is violated.

Suppose first that the business firms and individuals who receive bank loans decide to redeposit only a fraction of the proceeds into their bank accounts. Then, for example, the first $80,000 loan would lead to a deposit of less than $80,000—and similarly down the chain. The whole chain of deposit creation would therefore be reduced. Thus:

If individuals and business firms decide to hold more cash, the multiple expansion of bank deposits will be curtailed because fewer dollars of cash will be available to be used as reserves to support new checking deposits. Consequently, the money supply will be smaller.

The basic idea here is simple. Each $1 of cash held by a bank can support several dollars (specifically, $\$1/m$) of money. But $1 of cash held by an individual is exactly $1 of money; it supports no bank deposits. Hence, any time cash leaves the banking system, the money supply will decline. And any time cash enters the banking system, the money supply will rise.

Next, suppose that bank managers become more conservative, or that the outlook for loan repayments worsens because of a recession. Then banks might decide to keep more reserves than the legal requirement and lend out less than the amounts assumed in Figure 29-3. If this happens, banks further down the chain receive smaller deposits and, once again, the chain of deposit creation is curtailed. Thus:

If banks wish to keep excess reserves, the multiple expansion of bank deposits will be restricted. A given amount of cash will support a smaller supply of money than would be the case if banks held no excess reserves.

THE NEED FOR MONETARY POLICY

If we pursue this point a bit further, we will see why government regulation of the money supply is so important for economic stability. We have just suggested that banks will wish to keep excess reserves when they do not foresee profitable and secure opportunities to make loans. This is most likely to happen during the downswing and around the bottom of a business contraction. At such times, the propensity of banks to hold excess reserves can turn the deposit creation process into one of deposit destruction. Thus:

During a recession, profit-oriented banks would be prone to reduce the money supply by increasing their excess reserves—if the government did not intervene. As we will learn in subsequent chapters, the money supply is an important influence on aggregate demand, so such a contraction of the money supply would aggravate the recession.

This is precisely what happened—with a vengeance—during the Great Depression of the 1930s. Although total bank reserves grew, the money supply contracted violently because banks preferred to hold excess reserves rather than make loans that might not be repaid.

On the other hand, banks will want to squeeze the maximum possible money supply out of any given amount of cash reserves by keeping their reserves at the bare minimum when the demand for bank loans is buoyant, profits are high, and secure investment opportunities abound. This reduced incentive to hold excess reserves in prosperous times means that:

During an economic boom, the behavior of profit-oriented banks is likely to make the money supply expand, adding undesirable momentum to the booming economy and paving the way for a burst of inflation. The authorities must intervene to prevent this.

Regulation of the money supply, then, is necessary because profit-oriented bankers might otherwise provide the economy with a gyrating money supply that dances to the tune of the business cycle. Precisely how the authorities keep the money supply under control is the subject of the next chapter.

SUMMARY

1. It is more efficient to exchange goods and services by using **money** as a **medium of exchange** than by **bartering** them directly.
2. In addition to being the medium of exchange, whatever serves as money is likely to become the standard **unit of account** and a popular **store of value.**
3. Throughout history, all sorts of things have served as money. **Commodity money** gave way to full-bodied paper money (certificates backed 100 percent by some commodity, like gold), which in turn gave way to partially backed paper money. Nowadays our paper money has no commodity backing whatsoever; it is pure **fiat money.**
4. One popular definition of the U.S. money supply is **M1**, which includes coins, paper money, and several types of checking deposits. However, many economists prefer the **M2** definition, which adds to M1 other types of checkable accounts and most savings deposits. Much of M2 is held outside of banks.
5. Under our modern system of **fractional reserve banking,** banks keep cash reserves equal to only a fraction of their total deposit liabilities. This is the key to their profitability, since the remaining funds can be loaned out at interest. But it also leaves them potentially vulnerable to **runs.**

6. Because of this vulnerability, bank managers are generally conservative in their investment strategy. They also keep a prudent level of reserves. Even so, the government keeps a watchful eye over banking practices.
7. Before 1933, bank failures were common; but they declined sharply when **deposit insurance** was instituted.
8. Because it holds only fractional reserves, even a single bank can create money. But its ability to do so is severely limited because the funds it lends out probably will be deposited in another bank.
9. As a whole, the banking system can create several dollars of deposits for each dollar of reserves it receives. Under certain assumptions, the ratio of new deposits to new reserves will be $\$1/m$, where m is the required reserve ratio.
10. The same process works in reverse, as a system of money destruction, when cash is withdrawn from the banking system.
11. Because banks and individuals may want to hold more cash when the economy is shaky, the money supply would probably contract under such circumstances if the government did not intervene. Similarly, the money supply would probably expand rapidly in boom times if it were unregulated.

KEY TERMS

Run on a bank
Barter
Money
Medium of exchange
Unit of account
Store of value
Commodity money
Fiat money
M1
M2
Near moneys
Liquidity
Fractional reserve banking
Deposit insurance
Federal Deposit Insurance Corporation (FDIC)
Required reserves
Asset
Liability
Balance sheet
Net worth
Deposit creation
Excess reserves

QUESTIONS FOR REVIEW

1. If ours were a barter economy, how would you pay your tuition bill? What if your college did not want the goods or services you offered in payment?
2. How is "money" defined, both conceptually and in practice? Does the U.S. money supply consist of commodity money, full-bodied paper money, or fiat money?
3. What is fractional reserve banking, and why is it the key to bank profits? (*Hint:* What opportunities to make profits would banks have if reserve requirements were 100 percent?) Why does fractional reserve banking give bankers discretion over how large the money supply will be? Why does it make banks potentially vulnerable to runs?
4. During the 1980s and early 1990s, there was a rash of bank failures. Explain why these failures did not lead to runs on banks.
5. Suppose that no banks keep excess reserves and no individuals or firms hold on to cash. If someone suddenly discovers $12 million in buried treasure, explain what will happen to the money supply if the required reserve ratio is 10 percent.
6. How would your answer to Review Question 5 differ if the reserve ratio were 25 percent? If the reserve ratio were 100 percent?
7. Each year during Christmas shopping season, consumers and stores wish to increase their holdings of cash. Explain how this could lead to a multiple contraction of the money supply. (As a matter of fact, the authorities prevent this contraction from occurring by methods explained in the next chapter.)
8. Excess reserves make a bank less vulnerable to runs. Why, then, don't bankers like to hold excess reserves? What circumstances might persuade them that it would be advisable to hold excess reserves?
9. Use tables such as Tables 29-2 and 29-3 to illustrate what happens to bank balance sheets when each of the following transactions occurs:
   a. You withdraw $50 from your checking account to buy concert tickets.

b. Sam finds a $50 bill on the sidewalk and deposits it into his checking account.

c. Mary Q. Contrary withdraws $1,000 in cash from her account at Hometown Bank, carries it to the city, and deposits it into her account at Big City Bank.

10. For each of the transactions listed in Review Question 9, what will be the ultimate effect on the money supply if the required reserve ratio is 1/8 (12.5 percent)? Assume that the oversimplified deposit multiplier formula applies.

11. If the government takes over a failed bank with liabilities (mostly deposits) of $2.0 billion, pays off the depositors, and sells the assets for $1.5 billion, where does the missing $500 million come from? Why?

CHAPTER 30

MONETARY POLICY AND THE NATIONAL ECONOMY

Victorians heard with grave attention that the Bank Rate had been raised. They did not know what it meant. But they knew that it was an act of extreme wisdom.
J. K. Galbraith

Now that we understand the rudiments of the banking system, we are ready to bring money and interest rates into our model of income determination and the price level. In earlier chapters, we took investment (*I*) to be a fixed number. But this is a poor assumption. Not only is investment highly variable, it also depends on interest rates. And interest rates are, in turn, heavily influenced by *monetary policy.* The main task of this chapter is to explain how monetary policy affects interest rates, and thereby investment and aggregate demand.

We begin by learning about the operations of America's *central bank,* the *Federal Reserve System.* The "Fed," as it is called, is a very special kind of bank. Its customers are banks rather than individuals, and it performs some of the same services for them as your bank performs for you. Although it makes enormous profits, that is not its goal. Instead, the Fed tries to manage the money supply and interest rates in what it perceives to be the national interest. Just how the Fed does its job, and why its performance has fallen short of perfection, are the first subjects of this chapter.

Next, we integrate money and interest rates into our model of the macroeconomy. The mechanisms through which monetary policy affects aggregate demand are spelled out and analyzed in detail, and we learn an additional reason why the aggregate demand curve slopes downward. By the end of the chapter, we will have constructed a complete macroeconomic model, which we will use in the next few chapters to investigate a variety of important policy issues.

MONEY AND INCOME: THE IMPORTANT DIFFERENCE

But first we must get some terminology straight. The words *money* and *income* are used almost interchangeably in common parlance. This is a pitfall we must learn to avoid.

Money is a snapshot concept. It is the answer to questions like: "How much money do you have right now?" or "How much money did you have at 3:32 p.m. on Friday, November 5th?" To answer questions like these, you would add up the cash you are (or were) carrying and whatever checkable balances you have (or had), and answer something like: "I have $126.33," or "On Friday, November 5th, at 3:32 p.m., I had $31.43."

Income, by contrast, is more like a motion picture; it comes to you only over a period of time. If you are asked "What is your income?" you must respond by saying "$200 *per week*," or "$800 *per month*," or "$10,000 *per year*," or something like that. Notice that there is a unit of time attached to each of these responses. If you just say "My income is $452," without indicating whether it is per week or per month or per year, no one will understand what you mean.

That the two concepts are very different is easy to see. A typical American family has an *income* of perhaps $38,000 per year, but its holdings of *money* at any point in time (using the M1 definition) may be under $2,000. Similarly, at the national level, nominal GDP in 1995 was just over $7,200 billion, while the money stock (M1) in the middle of the year was only about $1,150 billion.

While money and income are very different, they are certainly related. This chapter is precisely about that relationship. Specifically, we will look at how the stock of *money* in existence at any moment of time influences the rate at which people will be earning *income*, that is, how money affects the GDP.

THE FEDERAL RESERVE SYSTEM: ORIGINS AND STRUCTURE

A **CENTRAL BANK** is a bank for banks. America's central bank is the *Federal Reserve System*.

When the *Federal Reserve System* was established in 1914, the United States joined the company of most of the other advanced industrial nations. Up until then, the United States, distrustful of centralization of economic power, was almost the only important nation without a **central bank**. Britain's central bank, the Bank of England, for example, dates from 1694.

The impetus for the establishment of a central bank in the United States came not from the power of economic logic but from some painful experiences with economic reality. Four severe banking panics between 1873 and 1907, in which many banks failed, convinced legislators and bankers alike that a central bank that would regulate credit conditions was not a luxury but a necessity. The 1907 crisis led to a study of the shortcomings of the banking system and, eventually, to the establishment of the Federal Reserve System.

Although the basic idea of central banking came from Europe, some changes were made when it was imported, making the Federal Reserve System a uniquely American institution. Owing to the vastness of our country, the extraordinarily large number of commercial banks, and our tradition of dual state-federal regulations, it was decided that the United States should have not one central bank but 12.

Technically, each of the Federal Reserve banks is a corporation; its stockholders are the banks that belong to it. But your bank, if it is a member of the system, does not enjoy the privileges normally accorded to stockholders: it receives only a token share of the Federal Reserve's immense profits (the bulk is turned over to the U.S. Treasury), and it has virtually no say in the decisions of the corporation. The private banks are more like customers of the Fed than like owners.

Who, then, controls the Fed? Most of the power resides in the seven-member Board of Governors of the Federal Reserve System in Washington, and especially

in its chairman, who is now Alan Greenspan, an economist. Members of the board are appointed by the president of the United States, with the advice and consent of the Senate, for 14-year terms. The president also designates one of the members to serve a 4-year term as chairman of the board, and thus to be the most powerful central banker in the world, for the Federal Reserve is *independent* of the rest of the government. As long as it stays within the statutory authority delineated by Congress, it alone has responsibility for determining the nation's monetary policy. The power of appointment, however, gives the president considerable long-run influence over Federal Reserve policy.

Closely allied with the Board of Governors is the powerful *Federal Open Market Committee (FOMC)*, which meets eight times a year in Washington. For reasons to be explained shortly, the decisions of the FOMC largely determine short-term interest rates and the size of the U.S. money supply. This 12-member committee consists of the seven governors of the Federal Reserve System, the president of the Federal Reserve Bank of New York, and, on a rotating basis, four of the other 11 district bank presidents.[1]

THE INDEPENDENCE OF THE FED

The institutional independence of the Federal Reserve System is looked upon as a source of pride by some and as an antidemocratic embarrassment by others.

Proponents of Federal Reserve independence argue that it enables monetary policy decisions to be made on objective, technical criteria and keeps monetary control out of the "political thicket." Without this independence, they argue, politicians might try to force the Fed to expand the money supply too rapidly, thereby contributing to chronic inflation and

be done. While full examination of the complexities of such stabilization policy requires more advanced discussion, the basic principles are simple enough to be grasped by beginning students—as we hope you have done.

But some economists argue that these lessons are best forgotten. In practice, they claim, attempts at macroeconomic stabilization are likely to do more harm than good. Policymakers are therefore best advised to follow fixed rules rather than use their best judgment on a case-by-case basis.

Nothing we have said so far leads to this conclusion. But we have not yet told the whole story. By the end of the chapter you will have encountered several arguments in favor of rules, and hence be in a better position to make up your own mind on this important debate.

VELOCITY AND THE QUANTITY THEORY OF MONEY

We saw in the last chapter how money influences real output and the price level in the Keynesian model. But there is another way to look at these matters, using a model that is much older than the Keynesian model. This model is known as the *quantity theory of money*, and it is easy to understand once we have introduced one new concept—*velocity*.

We learned in Chapter 29 that because barter is so cumbersome, virtually all economic transactions in advanced economies are conducted by the use of money. This means that if there are, say, \$5,000 billion worth of transactions in the economy during a particular year, and there is an average money stock of \$1,000 billion during that year, then each dollar of money is used an average of five times during the year.

VELOCITY indicates the number of times per year that an "average dollar" is spent on goods and services. It is the ratio of nominal GDP to the number of dollars in the money stock. That is:

$$\text{Velocity} = \frac{\text{Nominal GDP}}{\text{Money stock}}$$

The number 5 in this example is called the *velocity of circulation*, or just **velocity** for short, because it indicates the *speed* at which money circulates. For example, a particular dollar bill might be used to buy a haircut in January; the barber might use it to buy a sweater in March; the storekeeper might then use it to pay for gasoline in May; the gas station owner could pay it out to the house painter in October; and the painter might spend it on a Christmas present in December. This would mean that the dollar was used five times during the year. If it were used only four times during the year, its velocity would be 4, and so on. Similarly, a \$20 bill circulating with a velocity of 8 would be the monetary instrument used to finance \$160 worth of transactions in that year.

No one has data on all the transactions in the economy. To make velocity an operational concept, economists must settle on a workable definition of transactions that they can actually measure. The most popular choice is nominal gross domestic product, even though it ignores many transactions that use money—such as sales of existing assets. If we accept nominal GDP as a measure of the money value of transactions, we are led to a concrete definition of velocity as the ratio of nominal GDP to the number of dollars in the money stock. Since nominal GDP is the product of real GDP times the price level, we can write this definition in symbols as:

$$\text{Velocity} = \frac{\text{Value of transactions}}{\text{Money stock}} = \frac{\text{Nominal GDP}}{M} = \frac{P \times Y}{M}$$

The **EQUATION OF EXCHANGE** states that the money value of GDP transactions must be equal to the product of the average stock of money times velocity. That is:

$$M \times V = P \times Y$$

By multiplying both sides of the equation by M, we arrive at an identity called the **equation of exchange** that relates the money supply and nominal GDP:

$$\text{Money supply} \times \text{Velocity} = \text{Nominal GDP}$$

Alternatively, stated in symbols, we have:

$$M \times V = P \times Y$$

Here we have an obvious link between the stock of money, M, and the nominal value of the nation's output. But it is only a matter of arithmetic, not of economics. For example, it does not imply that the Fed can raise nominal GDP by increasing M. Why not? Because V might simultaneously fall by enough to prevent $M \times V$ from rising. In other words, if there were more dollar bills in circulation than before, but each bill changed hands more slowly, total spending might not rise. Thus, we need an auxiliary assumption to change the arithmetic into an economic theory:

The quantity theory of money transforms the equation of exchange from an accounting identity into an economic model by assuming that changes in velocity are so minor that velocity can be taken to be virtually constant.

You can see that if V never changed, the equation of exchange would be a marvelously simple model of the determination of nominal GDP—far simpler than the Keynesian model. To see this, it is convenient to rewrite the equation of exchange in growth-rate form:

$$\%\Delta M + \%\Delta V = \%\Delta P + \%\Delta Y$$

If V was constant (so its percentage change was always zero), this equation would say, for example, that if the Federal Reserve wanted to make nominal GDP grow by 8.7 percent per year, it need only raise the money supply by 8.7 percent per year. In such a simple world, economists could use the equation of exchange to *predict* nominal GDP growth simply by predicting the growth rate of money. And policymakers could *control* nominal GDP growth simply by controlling growth of the money supply.

In the real world things are not so simple because velocity is not a fixed number. But this does not necessarily destroy the usefulness of the quantity theory. We explained in Chapter 1 why all economic models make assumptions that are at least mildly unrealistic—without such assumptions they would not be models at all, just tedious descriptions of reality. The question is really whether the assumption of constant velocity is a useful abstraction from annoying detail or a gross distortion of facts.

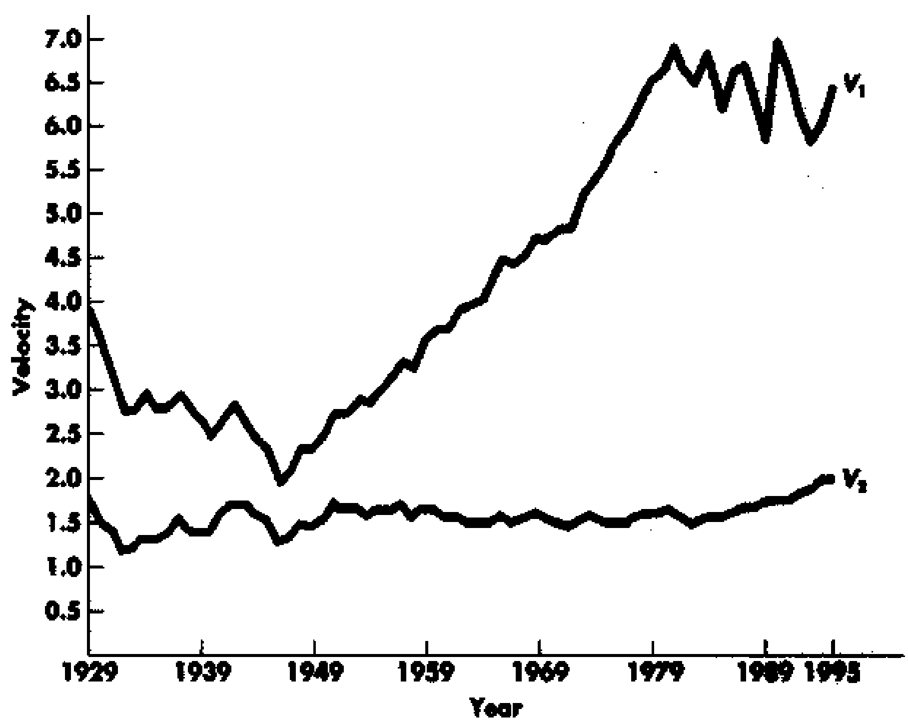
Figure 31-1 sheds some light on this question by showing the behavior of velocity since 1929. You will note that there are two different measures of velocity, labeled V_1 and V_2. Why? Recall from Chapter 29 that there are several ways to measure money, the most popular of which are M1 and M2. Since velocity (V) is simply nominal GDP divided by the money stock (M), we get a different measure of V for each measure of M. Figure 31-1 shows the velocities of both M1 and M2.

Several features are apparent. You will undoubtedly notice the difference in the behavior of V_1 versus V_2. There is a clear downward trend in V_1 from 1929 until 1946, a pronounced upward trend until 1981, and quite erratic behavior with some downward trend in recent years. Clearly, *the velocity of M1 is not constant over long periods of time. The velocity of M2 is much closer to constant,* but it has risen noticeably in recent years. Furthermore, closer examination of monthly or quarterly data on either V_1 or V_2 reveals some rather substantial fluctuations of velocity. Such fluctuations have led most economists to conclude that *velocity is not constant in the short run.* And predictions of nominal GDP growth based on

FIGURE 31-1 VELOCITY OF CIRCULATION, 1929–1995

During a period of American history of over 65 years, the velocity of M1 fell from almost 4 in 1930 to 2 in 1946, and then rose to almost 7 in 1981. Since then, it has gyrated dramatically. Clearly, then, M1 velocity has not been constant over long periods of time. The velocity of M2 looks much more stable, but it too has displayed some marked gyrations.

SOURCE: Constructed by the authors; data from Bureau of Economic Analysis, Federal Reserve Board, and Professor Robert Rasche.



assuming constant velocity have not fared very well, regardless of how M is measured. It seems, then, that the strict quantity theory of money is not an adequate model of aggregate demand.

THE DETERMINANTS OF VELOCITY

Since it is abundantly clear that velocity is a variable, not a constant, the equation of exchange is useful as a model of GDP determination only if we can explain movements in velocity. What factors decide whether V_1 or V_2 will be 4 or 5 or 6; that is, whether a dollar will be used to buy goods and services four or five or six times a year?

Perhaps the principal factor is the *frequency with which paychecks are received.* The idea can best be explained through a numerical example. Consider a worker who earns $24,000 a year, paid to her in 12 monthly paychecks of $2,000 each. Suppose she spends the whole $2,000 over the course of each month and maintains a minimum balance in the checking account of $500. Each payday her bank balance will jump to $2,500 and then gradually decline as she makes withdrawals to purchase goods and services. Finally, on the day before her next paycheck arrives, her checking balance will be just $500. Over the course of a typical month, then, her average checking account balance will be $1,500 (halfway between $2,500 and $500).

Now suppose her employer switches to a twice-a-month payroll. Her paychecks come twice as often, but are reduced to $1,000 each. There is no reason for her rate of spending to change, but her average *cash balance* will change. For now her checking balance will rise only to $1,500 on payday (the $500 minimum balance plus the $1,000 paycheck), and it will still be drawn down gradually to

$500. Her average cash balance will therefore decline to $1,000 (halfway between $1,500 and $500). Why is this so? Because, with the next paycheck coming sooner than before, it is not necessary to keep as much cash in the bank in order to carry out a given quantity of transactions.

But what does this have to do with velocity? Notice that when she was on a monthly payroll, this worker's personal velocity was:

$$V = \frac{\text{Annual income}}{\text{Average cash balance}} = \frac{\$24{,}000}{\$1{,}500} = 16$$

When she switched to a semimonthly payroll, velocity rose to:

$$V = \frac{\text{Annual income}}{\text{Average cash balance}} = \frac{\$24{,}000}{\$1{,}000} = 24$$

The general lesson to be learned is that:

More frequent wage payments mean that people can conduct their transactions with lower average cash balances. Since they will want to hold less cash, money will circulate faster. In other words, velocity will rise.

A second factor influencing velocity is the nature and *efficiency of the payments mechanism*, including how quickly checks clear banks, the use of credit cards, and other methods of transferring funds. It is easy to see how this works.

Our example assumed that the worker holds her entire paycheck in her checking account until she uses it to make a purchase. But, given that ordinary checking accounts pay little or no interest, this may not be sensible behavior. If it is possible to convert interest-bearing assets into money on short notice and at low cost, a rational individual might use her paycheck to purchase such assets and then use credit cards for most purchases, making periodic transfers to her checking account as necessary. For the same amount of total transactions, then, she would require lower money balances. This means that money would circulate faster: Velocity would rise.

The incentive to limit cash holdings depends on the ease and speed with which it is possible to exchange money for other assets. This is what we mean by the "efficiency of the payments mechanism." As computerization has speeded up the bookkeeping procedures of banks, as financial innovations have made it possible to transfer funds rapidly between checking accounts and other assets, and as credit cards have come to be used instead of cash, the need to hold money balances has declined. By definition, then, velocity has risen.

In practice, changes in the payments mechanism have posed severe practical problems for analysts interested in predicting velocity. A host of financial innovations, beginning in the 1970s and continuing to this day (some of which were mentioned in Chapter 29's discussion of the definitions of money), have made forecasting velocity a hazardous occupation. In fact, many economists believe the task is impossible and should not even be attempted.

A third determinant of velocity is the *rate of interest*. The basic motive for economizing on money holdings is that most money (at least M1) pays little or no interest, while many alternative stores of value pay higher rates. The higher these alternative rates of interest, the greater the incentive to economize on holding money. Therefore, as interest rates rise, people want to hold less money. So the existing stock of money circulates faster, and velocity rises.

It is this factor that most directly undercuts the usefulness of the quantity theory of money as a guide for monetary policy. For in the last chapter we learned

that expansionary monetary policy, which increases M, normally also decreases the interest rate. But if interest rates fall, other things being equal, velocity (V) will also fall. Thus, *when the Fed raises the money supply* (M), *the product* $M \times V$ *may go up by a smaller percentage than does* M *itself.*

One component of the interest rate is worth singling out for special attention: *the expected rate of inflation.* We explained in Chapter 23 why an "inflation premium" equal to the expected inflation rate often gets built into market interest rates.[1] Thus, in many instances, high inflation is the principal cause of high nominal interest rates. High rates of inflation, which erode the purchasing power of money, therefore lead both individuals and businesses to hold as little money as they can get by on—actions that increase velocity. To summarize this discussion of the determinants of velocity:

Velocity is not a strict constant but depends on such things as the frequency of payments, the efficiency of the financial system, the rate of interest, and the rate of inflation. Only by studying these determinants of velocity can we hope to predict the growth rate of nominal GDP from knowledge of the growth rate of the money supply.

MONETARISM: THE QUANTITY THEORY MODERNIZED

A group of economists called *monetarists* tries to do precisely this. Monetarists recognize that velocity changes, but they believe that such changes are fairly *predictable*—certainly in the long run and perhaps even in the short run. So they conclude that the best way to study economic activity is to start with the equation of exchange in growth-rate form:

$$\%\Delta M + \%\Delta V = \%\Delta P + \%\Delta Y$$

From here, careful study of the determinants of money growth (which we provided in the previous two chapters) and of changes in velocity (which we just completed) can be used to *predict* the growth rate of nominal GDP. Similarly, given an understanding of movements in V, control over M gives the Fed *control* over nominal GDP.

MONETARISM is a mode of analysis that uses the equation of exchange to organize and analyze macroeconomic data.

These are the central tenets of **monetarism.** When something happens in the economy, monetarists ask two questions:

1. What does this event do to the growth rate of money?
2. What does this event do to velocity?

From the answers, they assert that they can predict how the growth rate of nominal GDP will be affected.

By comparing the monetarist approach with the Keynesian approach that we described in the previous chapter, we can put both theories into perspective and understand the limitations of each. As we mentioned earlier, they differ more in style than in substance. Keynesians divide economic knowledge into four neat compartments—marked C, I, G, and $(X - IM)$—and unite them all with the equi-

[1] If you need review, turn back to pages 550–551.

librium condition that $Y = C + I + G + (X - IM)$. In Keynesian analysis, money affects the economy by first affecting interest rates.

Monetarists, on the other hand, organize their knowledge into two alternative boxes—labeled M and V—and then use a simple identity that says $M \times V = P \times Y$ to bring this knowledge to bear in predicting aggregate demand. In the monetarist model, the role of money in the national economy is not necessarily limited to working through interest rates.

The bit of arithmetic that multiplies M and V to get $P \times Y$ is neither more nor less profound than the one that adds up C, I, G and $(X - IM)$ to get Y. And certainly both are correct. The only substantive difference is that the monetarist equation leads to a prediction of *nominal* GDP while the Keynesian equation leads to a prediction of *real* GDP.

Why, then, do we not simply mesh the two theories—using the monetarist approach to study nominal GDP and the Keynesian approach to study real GDP? It seems that by doing so we could use the separate analyses of real and nominal GDP to obtain a prediction of the future behavior of the price level, which, of course, is the ratio of nominal GDP to real GDP.

The reason that this appealing procedure will not work helps point out the major limitation of each theory. *Taken by itself, either theory is incomplete.* Each gives us a picture of the *demand* side of the economy without saying anything about the *supply* side. To try to predict both the price level and real output solely from these demand-oriented models would be like trying to predict the price of peanuts by studying only the behavior of consumers and ignoring that of farmers. It just will not work. In terms of our earlier aggregate supply and demand analysis:

Both the monetarist and Keynesian analyses are ways of studying the *aggregate demand curve*. In neither case is it possible to learn anything about both output and the price level without also studying the *aggregate supply curve*.

Economists thus are forced to choose between two alternative ways of predicting aggregate demand. Those who choose the monetarist route will use velocity and the money supply to study aggregate demand in nominal terms. But then they must turn to the supply side to estimate how any predicted change in nominal demand gets apportioned between changes in production and changes in prices. The schematic diagram on page 712, with its emphasis on interest rates, plays little role in the monetarist analysis of how monetary policy affects the economy.

On the other hand, an economist working with the Keynesian approach will start by using the schematic diagram on page 712 to predict how monetary policy affects aggregate demand in real terms, that is, real GDP. But then he will have to turn to the aggregate supply curve to estimate the inflationary consequences of this real demand.

Which approach works better? There is no generally correct answer for all economies in all periods of time. When velocity behaved predictably in the 1960s and early 1970s, monetarism won many converts—in the United States and around the world. But, since then, velocity has behaved so erratically here and in many other countries that most economists have abandoned monetarism. Among the major industrial nations, only Germany still pays much attention to monetarist principles—and even there, erratic fluctuations in velocity have given the central bank fits.

FISCAL POLICY, INTEREST RATES, AND VELOCITY

We have now almost reconciled the Keynesian and monetarist views of how the economy operates. Because G is a part of $C + I + G + (X - IM)$, Keynesian analysis lends itself naturally to the study of fiscal policy. But we learned in the previous chapter that Keynesian economics also provides a powerful and important role for monetary policy: An increase in the money supply reduces interest rates, which, in turn, stimulates the demand for investment.

Monetarist analysis provides an obvious and direct route by which monetary policy influences both output and prices. But can the monetarist approach also handle fiscal policy? It can, because fiscal policy has an important effect on the rate of interest. Let us see how this works.

What happens to real output and the price level following, say, a rise in government purchases of goods and services? We learned in Chapter 28 that both real GDP (Y) and the price level (P) rise, and so nominal GDP certainly rises. But Chapter 30's analysis of the demand for money taught us that rising nominal GDP pushes the demand curve for money outward to the right. With no change in the supply curve for money, the rate of interest must rise. So *expansionary fiscal policy raises interest rates.*

If the government uses its spending and taxing weapons in the opposite direction, the same process works in reverse. Falling output and (possibly) falling prices shift the demand curve for money inward to the left. With a fixed supply curve for money, equilibrium in the money market leads to a lower interest rate. Thus:

Monetary policy is not the only type of policy that affects interest rates. Fiscal policy also affects interest rates. Specifically, increases in government spending or tax cuts normally push interest rates up, whereas restrictive fiscal policies normally pull interest rates down.

The fact that fiscal policy affects interest rates gives it a role in the monetarist model despite the fact that the equation of exchange, $M \times V = P \times Y$, does not include either government spending or taxation among its variables. Any fiscal policy that a Keynesian would call "expansionary"—higher spending, lower taxes, and so on—pushes up the rate of interest. And rising interest rates push up velocity because people want to hold less money when the interest they can earn on alternative assets increases. So it is through the V term in $M \times V$ that fiscal policy does its work in the monetarist framework. The equation of exchange then implies that nominal GDP must rise when, say, government spending increases—even if M is fixed—because velocity is higher.

Conversely, restrictive fiscal policies like tax increases and expenditure cuts reduce the quantity of money demanded and lower interest rates. The consequent drop in velocity reduces income through the equation of exchange, because the money supply circulates more slowly.

The translation, then, is complete. The Keynesian story about how fiscal policy works can be phrased in the monetarist dialect. And the monetarist tale about monetary policy can be told with a Keynesian accent. Furthermore, both modes of analysis help only to explain the mysteries of aggregate *demand* and must be supplemented by an analysis of aggregate *supply* to be complete. We must conclude, then, that:

The differences between Keynesians and monetarists have been grossly exaggerated by the news media. Indeed, when it comes to matters of basic economic theory, there are hardly any differences at all.

The fact that changes in fiscal policy move interest rates up and down has two other important consequences that merit some discussion.

APPLICATION: THE MULTIPLIER FORMULA REVISITED

We have just noted that expansionary fiscal policy raises interest rates. And we know that higher interest rates deter private investment spending. So when the government raises the G component of $C + I + G + (X - IM)$, one notable side effect will be to reduce the I component. Consequently, total spending will rise by less than simple multiplier analysis might suggest. The fact that a surge in government demand (G) discourages some private demand (I) provides another reason why the oversimplified multiplier formula, $1/(1 - MPC)$, exaggerates the size of the multiplier:

Because a rise in G (or, for that matter, an autonomous rise in any component of total expenditure) pushes interest rates higher, and hence deters some investment spending, the increase in the sum $C + I + G + (X - IM)$ is smaller than what the oversimplified multiplier formula predicts.

Combining this observation with our previous analysis of the multiplier, we now have the following complete list of:

REASONS WHY THE OVERSIMPLIFIED MULTIPLIER FORMULA IS WRONG

1. It ignores variable imports, which reduce the size of the multiplier.
2. It ignores price-level changes, which reduce the size of the multiplier.
3. It ignores the income tax, which reduces the size of the multiplier.
4. It ignores the rising interest rates that accompany any autonomous increase in spending, which also reduce the size of the multiplier.

Notice that all four of these adjustments point in the same direction. No wonder the actual multiplier (estimated to be below 2 for the U.S. economy) is so much less than the oversimplified formula suggests.

APPLICATION: DEFICIT REDUCTION AND INVESTMENT

We will discuss the government budget deficit in greater detail in the next chapter. But one major argument for reducing the deficit is that lower deficits should lead to higher levels of private investment spending. It is now simple to understand why. The government reduces its budget deficit by engaging in contractionary fiscal policies—lower spending or higher taxes. But we have just seen that any such measure should reduce real interest rates. These lower real interest rates should spur investment spending.

DEBATE: SHOULD STABILIZATION POLICY RELY ON FISCAL OR MONETARY POLICY?

We have seen that the Keynesian and monetarist approaches are more like two different languages than two different theories. However, it is well-known that language influences attitudes in many subtle ways. For example, the Keynesian language biases things subtly toward thinking first about fiscal policy simply because G is a part of $C + I + G + (X - IM)$ while M works indirectly.

Monetarists, on the other hand, think first about the equation of exchange, $M \times V = P \times Y$, which makes the effect of money on aggregate demand apparent and direct.

Years ago, economists engaged in a spirited debate in which extreme monetarists claimed that fiscal policy was futile, while extreme Keynesians argued that monetary policy was useless. But substantial evidence accumulated against both extreme positions, and these arguments are rarely heard today.

Instead of arguing over which type of policy is more powerful, economists nowadays debate which type of medicine—fiscal or monetary—cures the patient more quickly. Up to now, we have ignored questions of timing and pretended that the authorities instantly noticed the need for stabilization policy, decided upon a course of action, and administered the appropriate medicine. In reality, each of these steps takes time.

First, delays in data collection mean that the latest macroeconomic data pertain to the economy of a few months ago. Second, one of the prices of democracy is that the government often takes a distressingly long time to decide what should be done, to muster the necessary political support, and to put its decisions into effect. Finally, our $7 trillion economy is a bit like a sleeping elephant that reacts rather sluggishly to moderate fiscal and monetary prods. As it turns out, these *lags in stabilization policy*, as they are called, play a pivotal role in the choice between fiscal and monetary policy. Here's why.

The main policy tool for manipulating consumer spending (C) is the personal income tax, and Chapter 24 documented why the fiscal policy planner can feel fairly confident that each $1 of tax reduction will lead to about 90 to 95 cents of additional spending *eventually*. But not all of this will happen at once.

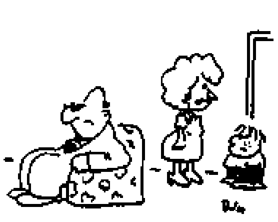
First, consumers must learn about the tax change. Then they may need to be convinced that the change is permanent. Finally, there is the simple force of habit: Households need time to adjust their spending habits when circumstances change. For all these reasons, consumers may increase their spending by only 30 to 50 cents for each $1 of additional income within the first few months after a tax cut. Only gradually will they raise their spending up to about 90 to 95 cents for each additional dollar of income.

Lags are much longer for investment (I), which provides the main vehicle by which monetary policy affects aggregate demand. Planning for capacity expansion in a large corporation is a long, drawn-out process. Ideas must be submitted and approved, plans must be drawn up, funding acquired, orders for machinery or contracts for new construction placed. And most of this occurs *before* any appreciable amount of money is spent. Economists have found that much of the response of investment to changes in interest rates or tax provisions is delayed for several *years*.

The fact that C responds more quickly than I has important implications for the choice among alternative stabilization policies. The reason is that the most common varieties of fiscal policy either affect aggregate demand directly—G is a component of $C + I + G + (X - IM)$—or work through consumption with a relatively short lag, while monetary policy has its major effects on investment. Therefore:

Conventional types of fiscal policy actions, such as changes in G or in personal taxes, probably affect aggregate demand much more promptly than do monetary policy actions.

This important fact was once used to argue that fiscal policy should bear the major burden of economic stabilization. But such a conclusion is a bit hasty, for



"Daddy's not mad at you, dear—Daddy's mad at the Fed."

the lags we have just described—which are beyond the control of policymakers—are not the only ones affecting the timing of stabilization policy. Further lags stem from the behavior of the policymakers themselves! We are referring here to the delays that occur while the policymakers study the state of the economy, contemplate what steps they should take, and finally put their decisions into effect. And here monetary policy has a huge advantage; that is:

Policy lags are normally much shorter for monetary policy than for fiscal policy.

The reasons are apparent. The Federal Open Market Committee (FOMC) meets eight times a year—and more often if necessary. So monetary policy decisions are made frequently. And once the Fed decides on a course of action, it is normally executed immediately by buying or selling bonds on the open market.

In contrast, federal budgeting procedures operate on an annual budget cycle. Except in rare circumstances, *major* fiscal policy initiatives can occur only at the time of the annual budget. In principle, tax laws can be changed at any time; but the wheels of Congress grind slowly and are often gummed up by partisan politics. So it may take many months for Congress to change fiscal policy. Even President Clinton's first budget, which passed Congress in record time in 1993, took almost 6 months from introduction to enactment. And in 1995, Congress failed to pass any budget before the government's fiscal year began, leading to two partial shutdowns of the federal government. In sum, one has to be very optimistic to suppose that the government can take important fiscal policy actions on short notice.

So where does the combined effect of expenditure lags and policy lags leave us? With nothing conclusive, we are afraid—at least in principle. In practice, however, most observers of contemporary America find it unrealistic to imagine using fiscal policy rationally for stabilization purposes until both partisanship and the budget deficit are sharply reduced. For now, but not necessarily forever, monetary policy appears to be the only game in town.

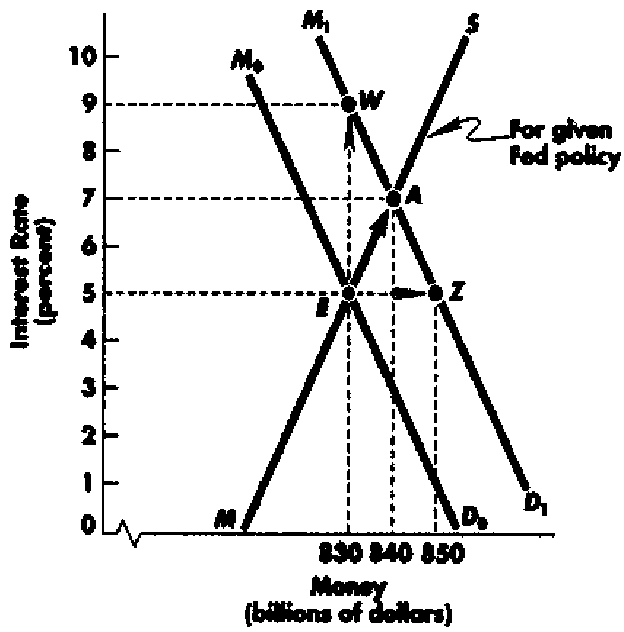
DEBATE: SHOULD THE FED CONTROL THE MONEY SUPPLY OR CONTROL INTEREST RATES?

Once we recognize that monetary policy must bear the current burden of stabilization policy, other questions arise. One major controversy, which raged for decades, was over how the Federal Reserve should conduct monetary policy. Some economists argued that the Fed should use its open-market operations to control the rate of interest (r) while others, especially monetarists, insisted that the Fed should concentrate on controlling some measure of the money supply (M). To understand the nature of this debate, we must first understand why the Fed cannot control both M and r at the same time.

Figure 31-2 will help us see why. It shows an initial equilibrium in the money market at point E, where money demand curve M_0D_0 crosses money supply curve MS. Here the interest rate is $r = 5$ percent and the money stock is $M =$ \$830 billion. Let us assume that these are the Fed's targets for M and r; it wants to keep the money supply and interest rates just where they are.

If the demand curve for money holds still, this is possible. But suppose the demand for money is not so obliging. Suppose, instead, that the demand curve shifts outward to the position indicated by M_1D_1 in Figure 31-2. We learned in the last chapter that this might happen because output increases or because prices rise. Or it might happen simply because people decide to hold more

FIGURE 31-2 THE FEDERAL RESERVE'S POLICY DILEMMA



This diagram illustrates the dilemma facing the Fed when the demand schedule for money shifts. In this case, we suppose that it increases from M_0D_0 to M_1D_1. If the equilibrium at point E satisfied its goals both for the money supply ($830 billion) and for the rate of interest (5 percent), either one or both of these goals will have to be abandoned after the demand schedule shifts. Points W, A, and Z illustrate three of the many choices. At W, the Fed is keeping the money supply at $830 billion through contractionary policies, but at a cost of skyrocketing interest rates. At Z, the Fed is holding interest rates at 5 percent, but the required expansionary monetary policies raise the money supply to $850 billion. At A, the Fed is not adjusting its policy and is accepting an increase in both the money supply and the rate of interest.

money. Whatever the reason, the Fed can no longer achieve both of its previous targets.

If the Fed takes no action, the outward shift in the demand curve will push up both the quantity of money (M) and the rate of interest (r). Figure 31-2 shows this graphically as the move from point E to point A. If the demand curve for money shifts outward from M_0D_0 to M_1D_1, and there is no change in monetary policy (so the supply schedule does not move), the money stock rises to $840 billion and the interest rate rises to 7 percent.

But suppose, first, that the Fed is targeting the money supply and is therefore unwilling to let M rise. In that case, it must use contractionary open-market operations to prevent M from rising. But, in so doing, it will push r up even higher—as Figure 31-2 shows. After the demand curve for money shifts out, point E is unattainable. The Fed must choose from among the points on M_1D_1, and point W is the point on this curve that keeps the money supply at $830 billion. By pushing the supply curve inward so that it passes through point W, the Fed can hold M at $830 billion. But the interest rate will skyrocket to 9 percent.

Alternatively, if the Fed is pursuing an interest rate target, it might decide that a rise in r is to be avoided. In this case, the Fed would be forced to engage in expansionary open-market operations to prevent the outward shift of the demand curve for money from pushing r up. In terms of Figure 31-2, the interest rate can be held at 5 percent by shifting the supply curve outward to pass through point Z. But to do this, the Fed will have to let the money supply rise to $850 billion. To summarize this discussion:

When the demand curve for money shifts outward, the Fed must tolerate a rise in interest rates, a rise in the money stock, or both. It simply does not have the weapons to control *both* the supply of money *and* the interest rate. If it tries to keep M steady, then r will rise sharply. Conversely, if it tries to stabilize r, then M will shoot up.

TWO IMPERFECT ALTERNATIVES

For years, economists have debated how the Fed should deal with its inability to control both the money supply and the rate of interest. Should it adhere rigidly to its target growth path for the money supply, regardless of the consequences for interest rates? Should it hold interest rates steady even if that causes wild gyrations in the money stock? Or is some middle ground more appropriate? Let us explore the issues before considering what has actually been done.

The main problem with rigid targets for the *supply* of money is that the *demand* for money does not cooperate by growing smoothly and predictably from month to month; instead it dances about quite a bit in the short run. This confronts the recommendation to control the money supply with two problems:

1. It is almost impossible to achieve. Since the volume of money in existence depends on *both* the demand *and* supply curves, keeping M on target in the face of significant fluctuations in demand for money would require exceptional dexterity on the part of the Fed.
2. For reasons just explained, rigid adherence to money-stock targets might lead to wide fluctuations in interest rates, which could create an unsettled atmosphere for business decisions.

By the same token, even more powerful objections can be raised against exclusive concentration on interest rate movements. Since increases in nominal GDP shift the demand schedule for money outward (as in Figure 31-2), a central bank determined to keep interest rates from rising would have to expand the money supply in response. Conversely, when GDP sagged, it would have to contract the money supply to keep rates from falling. Thus, interest rate *pegging* would make the money supply expand in boom times and contract in recessions—with potentially grave consequences for the stability of the economy. Ironically, this is precisely the sort of monetary behavior the Federal Reserve System was designed to prevent. Hence, if the Fed is to control interest rates, it had better formulate flexible targets, not fixed ones.

WHAT HAS THE FED ACTUALLY DONE?

In the early part of the postwar period, the predominant Keynesian view held that the interest rate target was much the more important of the two. The rationale was that gyrating interest rates would cause abrupt and unsettling changes in investment spending, which in turn would make the whole economy fluctuate. Stabilizing interest rates was therefore believed to be the best way to stabilize GDP. If doing so required fluctuations in the money supply, so be it. Consequently, the Fed focused on interest rates and paid little attention to the money supply.

In the 1960s, this prevailing view came under attack by Professor Milton Friedman and other monetarists. They argued that the Fed's obsession with stabilizing interest rates actually *destabilized* the economy by making the money supply fluctuate too much. The monetarist prescription was simple: The Fed should stop worrying about fluctuations in interest rates and make the money supply grow at a constant rate from month to month and year to year.

Monetarism made important inroads at the Fed during the inflationary 1970s. Early in the decade, the central bank began to keep much closer tabs on the money stock than it previously had. More important, a major change in the conduct of monetary policy was announced by then-Chairman Paul Volcker in

October 1979. Henceforth, he asserted, the Fed would stick more closely to its target for money-stock growth regardless of the implications for interest rates. Interest rates would go wherever the law of supply and demand took them.

According to our analysis, this change in policy should have led to wider fluctuations in interest rates. And it did. Unfortunately, the Fed ran into some bad luck. The ensuing 3 years were marked by unusually severe gyrations in the demand for money, so the ups and downs of interest rates were far more extreme than anyone had expected. Figure 31-3 gives an indication of just how volatile interest rates were between late 1979 and late 1982. Naturally, this erratic performance provoked some heavy criticism of the Fed.

Then, in October 1982, Chairman Volcker announced that the Fed was temporarily abandoning its attempts to stick to a target growth path for the money supply. Although he did not say so, his announcement presumably meant that the Fed started once again to pay more attention to interest rate targets. As you can see in Figure 31-3, interest rates became much more stable after the change in policy. Most observers think this was no coincidence.

After 1982, the Fed gradually distanced itself from the position that the money supply should grow at a constant rate. Finally, in 1993, Chairman Alan Greenspan officially confirmed what was already widely known: that the Fed was no longer using the various *M*s to guide policy. He strongly hinted that the Fed was targeting interest rates, especially *real* interest rates, instead—a hint that was repeated several times in 1994 and 1995. In truth, the Fed had little choice. The demand curve for money behaved so erratically and so unpredictably in the 1980s and 1990s that stabilizing the money stock was probably impossible and certainly undesirable. Whether this situation will continue is anyone's guess. But as of this writing, the Fed has shown no inclination to return to the *M*s.

DEBATE: THE SHAPE OF THE AGGREGATE SUPPLY CURVE

Another lively debate over stabilization policy revolves around the shape of the economy's aggregate supply curve. Many economists think of the aggregate supply curve as quite flat—as in Figure 31-4(a)—so that large increases in output

FIGURE 31-3 THE BEHAVIOR OF INTEREST RATES, 1979–1985

This chart traces interest rate movements from 1979 through 1985. Notice the extreme volatility of rates during the period from late 1979 to mid-1982—the period in which the Fed was concentrating more on stabilizing the money supply. After mid-1982, interest rates became much less volatile.

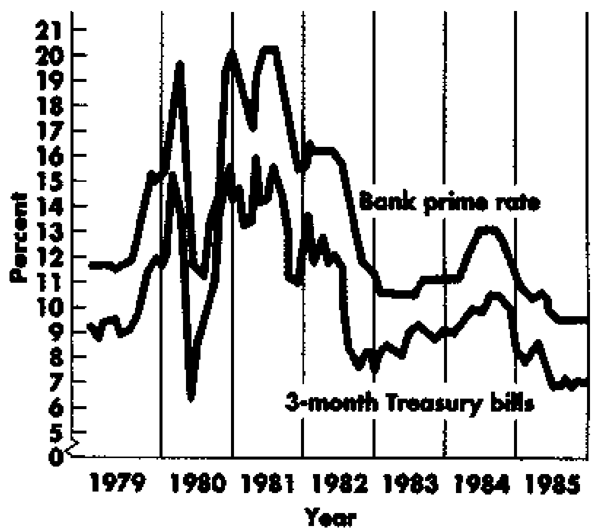
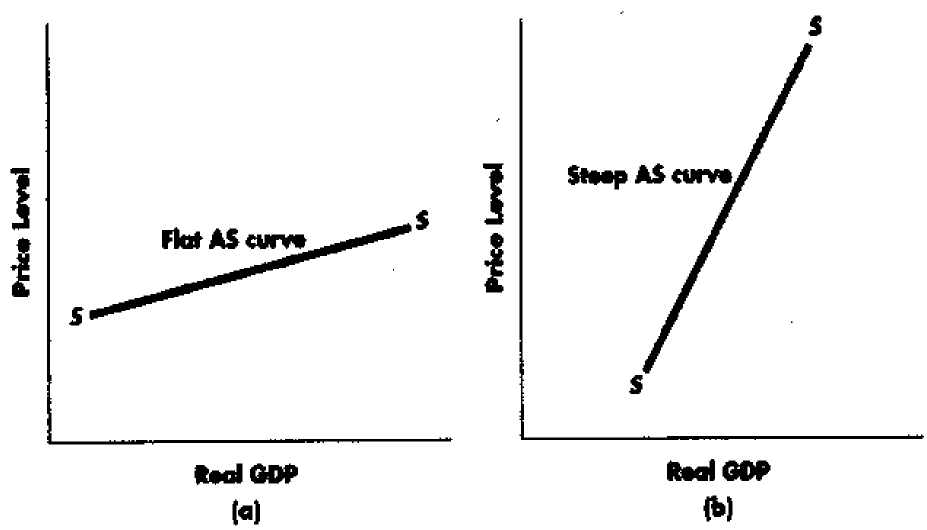


FIGURE 31-4

ALTERNATIVE VIEWS OF THE AGGREGATE SUPPLY CURVE

Some economists think of the economy's aggregate supply schedule as very flat, as in Panel (a). Others think of it as quite steep, as in Panel (b).



can be achieved with little inflation. But other economists envision the supply curve as steep, as in Figure 31-4(b), so that prices are very responsive to changes in output. The differences for public policy are substantial.

If the aggregate supply curve is flat, expansionary fiscal or monetary policy that raises the aggregate demand curve can buy large gains in real GDP at low cost in terms of inflation. In Figure 31-5(a), stimulation of demand pushes the aggregate demand curve outward from D_0D_0 to D_1D_1, thereby moving the economy's equilibrium from point E to point A. There is a substantial rise in output ($400 billion) with only a pinch of inflation (1 percent).

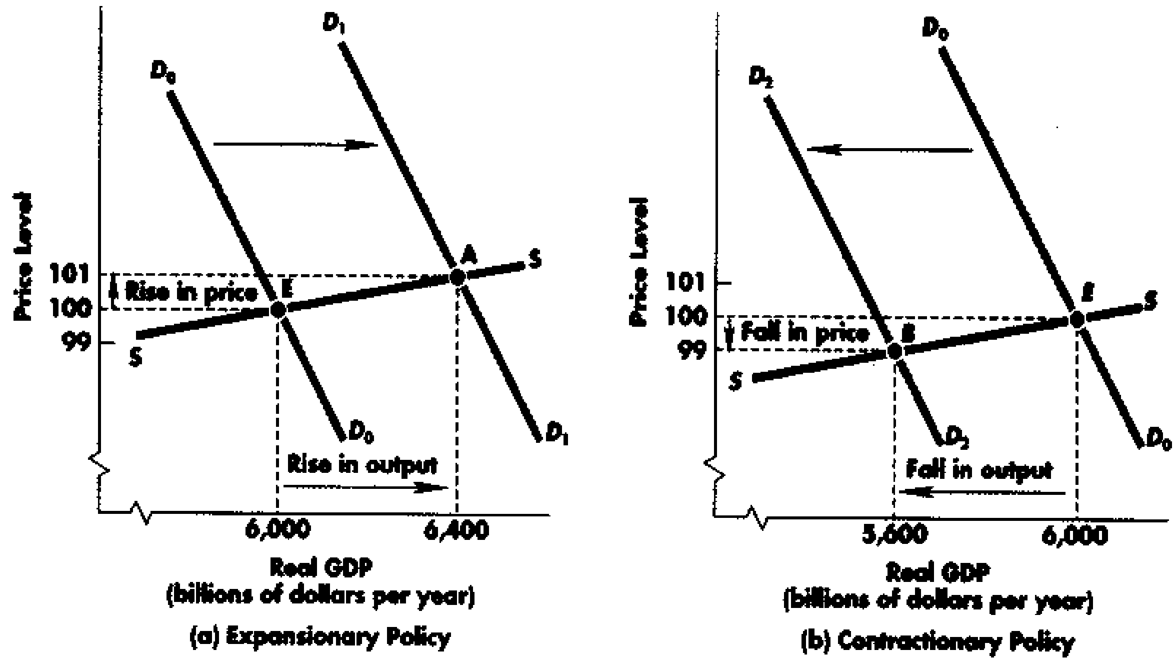
Conversely, when the supply curve is flat, a restrictive stabilization policy is not a very effective way to bring inflation down; instead, it serves mainly to reduce real output, as Figure 31-5(b) shows. Here, a leftward shift of the aggregate demand curve moves equilibrium from point E to point B, lowering real GDP by $400 billion but cutting the price level by merely 1 percent.

Things are quite different if the aggregate supply curve is steep. In that case, expansionary fiscal or monetary policies will cause a good deal of inflation without adding much to real GDP. [See Figure 31-6(a) on page 733, where expansionary policies shift equilibrium from E to A.] Similarly, contractionary policies are effective ways of bringing down the price level without much sacrifice of output, as shown by the shift from E to B in Figure 31-6(b).

The resolution of this debate is of fundamental importance for the proper conduct of stabilization policy. If the supply curve is flat, stabilization policy is much more effective at combating recession than inflation. If the supply curve is steep, precisely the reverse is true.

But why does the argument persist? Why cannot economists determine the shape of the aggregate supply curve and stop arguing? The answer is that supply conditions in the real world are far more complicated than our simple diagrams suggest. Some industries may have flat supply curves while others have steep ones. For reasons explained in Chapter 27, supply curves shift over time. And, unlike many laboratory scientists, economists cannot perform the controlled experiments that would reveal the shape of the aggregate supply curve directly. Instead, they must use statistical inference to make educated guesses.

FIGURE 31-5 STABILIZATION POLICY WITH A FLAT AGGREGATE SUPPLY CURVE



These two diagrams show that stabilization policy is much more effective as an antirecession policy than as an anti-inflation policy when the aggregate supply curve is flat. In Panel (a), monetary or fiscal policies push the aggregate demand curve outward from D_0D_0 to D_1D_1, causing equilibrium to shift from point E to point A. You can see that output rises substantially (from \$6,000 billion to \$6,400 billion), while prices rise only slightly (from 100 to 101, or 1 percent). So the policy is quite successful. In Panel (b), contractionary policies are used to combat inflation by pushing the aggregate demand curve inward from D_0D_0 to D_2D_2. Prices do fall slightly (from 100 to 99) as equilibrium shifts from point E to point B, but real output falls much more dramatically (from \$6,000 billion to \$5,600 billion); so the policy has had little success.

Although empirical research is proceeding, our understanding of aggregate supply remains much less settled than our understanding of aggregate demand. Nevertheless, many economists believe that the outline of a consensus view has emerged. This view holds that the steepness of the aggregate supply schedule depends on the time period under consideration.

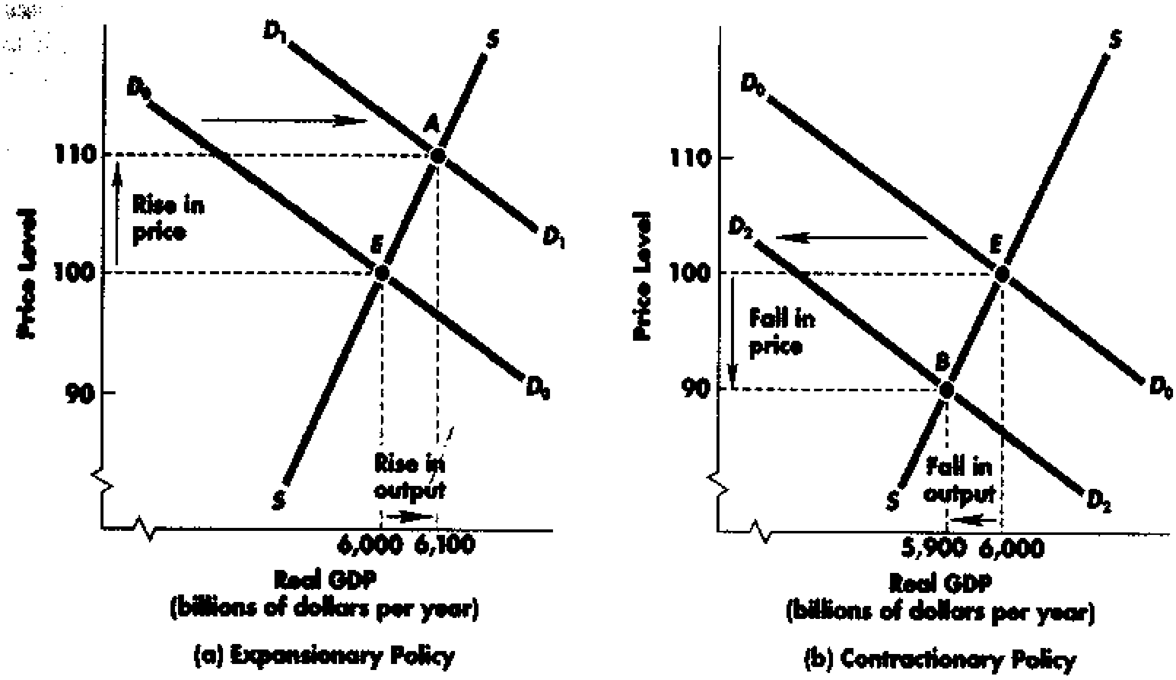
In the very short run, the aggregate supply curve is quite flat, so Figure 31-5 applies. Over short time periods, therefore, fluctuations in aggregate demand have large effects on output but only minor effects on prices. In the long run, however, the aggregate supply curve becomes quite steep, perhaps even vertical. In that case, Figure 31-6 applies, so that changes in demand affect mainly prices, not output.[2] The implication is that:

Any change in aggregate demand will have most of its effect on *output* in the short run but on *prices* in the long run.

Not all economists accept this middle-of-the-road view, but many do.

[2]The reasoning behind the view that the aggregate supply curve is flat in the short run but steep in the long run will be developed in Chapter 33.

FIGURE 31-6 STABILIZATION POLICY WITH A STEEP AGGREGATE SUPPLY CURVE



These two diagrams show that stabilization policy is much more effective at fighting inflation than at fighting recession when the aggregate supply curve is steep. In Panel (a), expansionary policies that push aggregate demand outward from D_0D_0 to D_1D_1 raise output by only \$100 billion but push up prices by 10 percent as equilibrium moves from point *E* to point *A*. So demand management is not a good way to end a recession. In Panel (b), contractionary policies that pull aggregate demand inward to D_2D_2 are successful in that they lower prices markedly (from 100 to 90, or about 10 percent) but reduce output only slightly (from \$6,000 billion to \$5,900 billion).

DEBATE: SHOULD THE GOVERNMENT INTERVENE?

We have yet to consider what may be the most fundamental and controversial debate of all—the issue posed as a puzzle at the beginning of the chapter. Is it likely that the government can conduct a successful stabilization policy? Or are its well-intentioned efforts likely to be harmful, so that it would be better to adhere to fixed rules?

This controversy has raged for several decades, and there is no end in sight. It is partly a political or philosophical debate because economists, like other people, come with both liberal and conservative stripes. Liberal economists tend to be more intervention-minded and hence look more favorably on discretionary stabilization policy. Conservative economists are more inclined to keep the government's hands off the economy and hence are more attracted to fixed rules. Such political differences are not surprising. But more than ideology propels the debate. We need to understand the economic issues.

Critics of stabilization policy point to the lags and uncertainties that surround the operation of both fiscal and monetary policies—lags and uncertainties that we have stressed repeatedly in this and earlier chapters. Will the Fed's actions have the desired effects on the money supply? What will these actions do to

interest rates and spending? Can fiscal policy actions be taken promptly? How large is the expenditure multiplier? The list could go on and on.

They look at this formidable catalogue of difficulties, add a dash of skepticism about our ability to forecast the future state of the economy (see below), and worry that stabilization policy may do more harm than good. These skeptics advise both the fiscal and monetary authorities to pursue a passive policy rather than an active one—adhering to fixed rules that, while incapable of ironing out every bump in the economy's growth path, will at least keep it roughly on track in the long run.

Advocates of active stabilization policy admit that perfection is unattainable. But they are much *more optimistic* about the prospects for success. And they are much *less optimistic* about how smoothly the economy would function in the absence of demand management. They therefore advocate discretionary increases in government spending (or decreases in taxes) and lower interest rates when the economy has a recessionary gap. Such policies, they believe, will help keep the economy closer to its full-employment growth path.

Naturally, each side can point to evidence that buttresses its own view. Activists look back with pride at the tax cut of 1964 and the sustained period of economic growth that it helped usher in. They also point to the tax cut of 1975, which was enacted at just about the trough of a severe recession, the Federal Reserve's switch to easy money in 1982, and the Fed's expert steering of the economy from 1992 to 1995. Advocates of rules remind us of the government's refusal to curb what was obviously a situation of runaway demand during the 1966–1968 Vietnam buildup, its overexpansion of the economy in 1972, the monetary overkill that helped bring on the sharp recession of 1981–1982, and the inadequate antirecession policies of the early 1990s.

The historical record of fiscal and monetary policy is far from glorious. It shows that while the authorities have sometimes taken appropriate and timely action to stabilize the economy, there are also many cases in which they either took inappropriate steps or did nothing at all. The question of whether the government should adopt passive rules or attempt an activist stabilization policy therefore merits a closer look. As we shall see, the lags in the effects of policy that we discussed earlier in this chapter play a pivotal role in the debate.

LAGS AND THE RULES-VERSUS-DISCRETION DEBATE

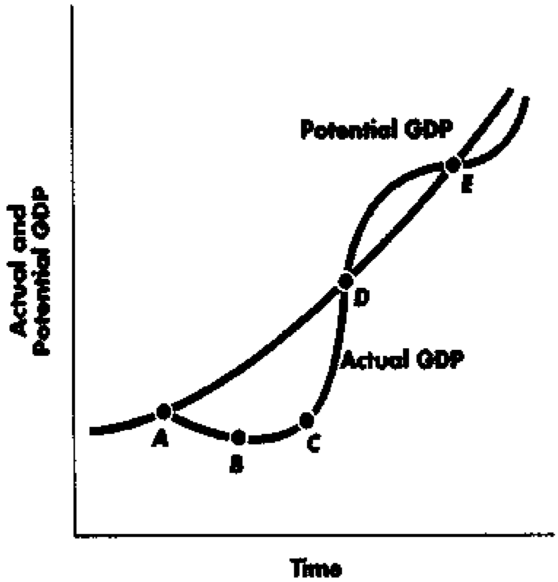
The reason why lags lead to a fundamental difficulty for stabilization policy—a difficulty so formidable that it has led some economists to conclude that attempts to stabilize economic activity are likely to do more harm than good—can be explained best by referring to Figure 31-7. Here we chart the behavior of both actual and potential GDP over the course of a business cycle in a hypothetical economy with no stabilization policy. At point *A*, the economy begins to slip into a recession and does not recover to full employment until point *D*. Then, between points *D* and *E*, it overshoots and is in an inflationary boom.

The case for stabilization policy runs like this. Policymakers recognize that the recession is a serious problem at point *B* and take appropriate actions. These actions have their major effects around point C and therefore curb both the depth and the length of the recession.

But suppose the lags are really longer and less predictable than this. Suppose, for example, that actions do not come until point C and that stimulative policies do not have their major effects until after point *D*. Then policy will be of little

FIGURE 31-7 A TYPICAL BUSINESS CYCLE

This is a stylized representation of the relationship between actual and potential GDP during a typical business cycle. The imaginary economy slips into a recession at point *A*, bottoms out around point *B*, and is in a recovery period until point *D*. After point *D*, it enters an inflationary boom that lasts until point *E*.



help during the recession and will actually do harm by overstimulating the economy during the ensuing boom. Thus:

In the presence of long lags, attempts at stabilizing the economy can actually succeed in destabilizing it.

Because of this, some economists argue that we are better off letting the economy alone and relying on its natural self-corrective forces to cure recessions and inflations. Instead of embarking on periodic programs of monetary and fiscal stimulus or restraint, they advise policymakers to stick to fixed rules, that is, to rigid formulas that ignore current economic events.

AUTOMATIC STABILIZERS

We have already mentioned the monetarist policy rule: The Fed should keep the money supply growing at a constant rate. For fiscal policy, proponents of rules often recommend that the government resist temptations to manage aggregate demand actively and rely instead on **automatic stabilizers**—features of the economy that reduce its sensitivity to shocks.

An **AUTOMATIC STABILIZER** is any arrangement that automatically serves to support aggregate demand when it would otherwise sag and to hold down aggregate demand when it would otherwise surge ahead. In this way, an automatic stabilizer reduces the sensitivity of the economy to shifts in demand.

Examples of automatic stabilizers are not hard to find in the federal budget. The personal income tax is the most obvious example. The tax acts as a shock absorber because it makes disposable income, and thus consumer spending, less sensitive to fluctuations in GDP. When GDP rises, disposable income (*DI*) rises less because part of the increase in GDP is siphoned off by the U.S. Treasury. This helps limit the upward fluctuation in consumption spending. And when GDP falls, *DI* falls less sharply because part of the loss is absorbed by the Treasury rather than by consumers. So consumption does not drop as much as it otherwise might. Thus, as we noted in Chapter 28, income taxes lower the value of the multiplier. In truth, the much-maligned personal income tax is one of several features of our modern economy that helps ensure against a repeat performance of the Great Depression.

There are many other automatic stabilizers in our economy. For example, Chapter 23 discussed the U.S. system of unemployment insurance. This serves as an automatic stabilizer in a similar way. When GDP drops and people lose their jobs, unemployment benefits prevent the disposable incomes of the jobless from falling as dramatically as their earnings. As a result, unemployed workers can maintain their spending, and consumption fluctuates less than employment.

And the list could continue. The basic principle is the same: Each automatic stabilizer serves, in one way or another, as a shock absorber, thereby lowering the multiplier. And each does so without the need for any decision maker to take action. Unlike discretionary policy changes, automatic stabilizers are not bothered by either long or unpredictable lags because, in a word, they work *automatically*.

DEBATE: RULES OR DISCRETION?

Believers in fixed rules assert that we should forget about discretionary policy and put the economy on automatic pilot—relying on automatic stabilizers and the economy's natural, self-correcting mechanisms. Are they right? As usual, the answer depends on many factors.

HOW FAST DOES THE ECONOMY'S SELF-CORRECTING MECHANISM WORK?

We emphasized in Chapter 27 that the economy has a self-correcting mechanism. If the economy can cure recessions and inflations quickly by itself, then the case for intervention is weak. For if such problems typically last only a short time, then lags in discretionary stabilization policy mean that the medicine will often have its major effects only after the disease is over. In terms of Figure 31-7, this would be a case where point D comes very close to point A.

While extreme advocates of rules argue that this is what indeed happens, most economists agree that the economy's self-correcting mechanism is slow and not terribly reliable, even when supplemented by the automatic stabilizers. On this count, then, a point is scored for discretionary policy.

HOW LONG ARE THE LAGS IN STABILIZATION POLICY?

We explained earlier why long and unpredictable lags in the formulation or operation of stabilization make it unlikely that policy can do much good. Short, reliable lags point in the opposite direction. Thus, advocates of fixed rules emphasize the length of lags while proponents of discretion discount them.

Who is right? It all depends on the circumstances. Sometimes fiscal policy actions are taken promptly, and the economy receives much of the stimulus from expansionary policy within a year after slipping into a recession. While far from perfect, such timely actions certainly would be felt soon enough to do some good. But, as we have seen, very slow fiscal responses may actually be destabilizing. Since history offers examples of each type, no general conclusion can be drawn.

HOW ACCURATE ARE ECONOMIC FORECASTS?

One way to cut down the policy-making lag enormously is to have good economic forecasts. If we could see a recession coming a full year ahead of time

(which we certainly *cannot* do), even a rather sluggish policy response would still be timely. In terms of Figure 31-7, this would be a case where the recession is predicted well before point *A*.

It therefore behooves us to take a look at the techniques that economists in universities, government agencies, and private businesses have developed over the years to assist them in predicting what the economy will do. There are a variety of techniques, none of them foolproof.

TECHNIQUES OF ECONOMIC FORECASTING

THE USE OF ECONOMETRIC MODELS

An **ECONOMETRIC MODEL** is a set of mathematical equations that embody the economist's model of the economy.

Among the most widely publicized forecasts are those generated by the use of **econometric models** of the economy. Put simply, an econometric model is merely a mathematical version of the models of macroeconomic activity that we have described in Parts VI and VII. The difference is that the basic notions are cast in equations rather than in diagrams. For example, our consumption function could have been expressed by the formula:

$$C = a + bDI$$

where C is consumer spending and DI is disposable income, instead of by a graph.[3]

The builder of an econometric model takes equations like these and uses actual data to estimate the sizes of a and b. For example, statistical analysis may lead to the conclusion that the correct magnitude of a in the previous formula is approximately 300 and that the most reasonable value of b is 0.75. Then the consumption function formula is:

$$C = 300 + 0.75DI$$

that is, consumer spending is $300 billion plus 75 percent of disposable income. The economist can complete the model by adding a definition of disposable income as GDP minus tax receipts:

$$DI = Y - T$$

and appending the fact that GDP is the sum of C, I, G and $(X - IM)$:

$$Y = C + I + G + (X - IM)$$

This simple model has a total of three equations. If we hypothesize that government purchases, tax receipts, net exports, and investment are all unaffected by the relationships in the model, then these three equations are just enough to determine the values of the three remaining variables: C, DI, and Y. These last three variables are called the model's *endogenous variables*, meaning that their values are determined *inside* the model. The remaining variables—G, T, I, and $(X - IM)$—are called the model's *exogenous variables* because they must be provided from *outside* the model. Given an econometric model, it takes only algebra to turn forecasts of the exogenous variables into corresponding forecasts of the endogenous variables.

Models actually used to forecast the behavior of the U.S. economy have hundreds of variables and equations. Because of their complexity, the only practical

[3]This is nothing but the formula for a straight line with a slope of b and an intercept of a.

way to solve them for forecasts of all the endogenous variables is to use a high-speed computer. But making the forecasts accurate is another thing entirely because of the "GIGO" (garbage in, garbage out) problem: If you feed garbage into a computer, that's exactly what will come out at the end. In the forecasting context, this "garbage" can be either inaccurate forecasts of the exogenous variables or bad equations. This is why model builders are constantly seeking to improve their equations. But they have yet to achieve perfection.

And, even if they did, their forecasts would still not be infallible because there is a certain amount of unavoidable randomness in macroeconomic behavior. After all, we are dealing with the behavior of literally millions of individuals and business firms, and events essentially outside our control can sometimes exert a profound influence on our economy. (Example: The Gulf war in 1990 made both inflation and unemployment higher than forecasters expected.) So forecasts of the exogenous variables are bound to be wide of the mark at times. Furthermore, econometric models are basically complicated statistical summaries of the past. No one can really be sure that the future will be like the past. But this is what we assume whenever we forecast with an econometric model.

LEADING INDICATORS

A **LEADING INDICATOR** is a variable that, experience has shown, normally turns down before recessions start and turns up before expansions begin.

A second forecasting method exploits observed historical timing relationships through the use of certain **leading indicators** that have in the past given advance warning of economic events.

For example, the stock market is a leading indicator because stock market downturns normally begin several months before downturns in industrial production. *Why* does this happen? Does the decline in the stock market cause economic downturns by reducing consumer spending? Or are both the stock market and industrial production just reacting to some other influence, with the stock market's reaction coming sooner? Certainly these are fascinating questions. But the answers may not be crucial to a forecaster *if* the stock market continues to be as good a leading indicator of industrial production in the future as it has been in the past. In that event, she can make use of the observed relationship between stock prices and industrial production for forecasting even if she does not entirely understand it.

As it turns out, however, excessive reliance on any single leading indicator produces an unimpressive forecasting record. An obvious solution is to look at many indicators. But when we do this, we often receive conflicting signals. If one indicator is rising rapidly while another is falling, what are we to do?

One way to resolve this conflict is to combine several leading indicators into one. For example, every month the news media report the latest reading on the Conference Board's composite index of 11 leading indicators. One frequently cited rule of thumb holds that three consecutive monthly declines in this index indicate that a recession is on the way. Unfortunately, history shows that this rule signals many downturns that never come. (The most recent example was in 1995.)

SURVEY DATA

A third method of forecasting utilizes periodic surveys of the intentions of business and consumers. These surveys are conducted by government agencies, by business groups like the Conference Board and the National Association of Purchasing Managers, and by the Survey Research Center of the University of Michigan. Their findings are published in the financial press and are widely used by

economists in industry, government, and academia. Some economists have found them useful in forecasting, say, consumer spending and industrial production.

JUDGMENTAL FORECASTS

This term is used to describe the forecasts of those opportunistic (and undoubtedly prudent!) forecasters who refuse to rely on any one method, but look instead at every available scrap of evidence. They study the outputs of the econometric models; they watch the leading indicators; and they scrutinize the findings of surveys. At times, it seems, they even gaze at the stars. Somehow, judgmental forecasters distill all this information in their heads and arrive at a forecast of GDP and other key variables. How do they go about it? An outside observer can never really tell, since the very nature of judgmental forecasting precludes the existence of a formula that can be written down or precisely described.

Official forecasts of the Federal Reserve, the Council of Economic Advisers, and the Congressional Budget Office are all judgmental. So are virtually all forecasts made by private business economists.

THE ACCURACY OF ECONOMIC FORECASTS

Which method wins the prize for the most accurate forecasts? First, no technique is clearly superior all the time. If it were, no one would use any other method. Second, because econometric forecasters use surveys, lead-lag patterns, and judgment in forming their predictions of exogenous variables, and since judgmental forecasters watch the models and surveys closely, a clean comparison is impossible. However, it seems that the most accurate forecasts have been derived by judgmental adjustment of forecasts from econometric models.

How accurate are economic forecasts? That depends both on the variable being forecast (consumption, for example, is easier than investment) and on the time period (for example, recession years are particularly difficult times for forecasters). To give a rough idea of magnitudes, forecasts of either the inflation rate or the real GDP growth rate for the year ahead typically err by plus or minus 3/4 to 1 percentage point. But, in a bad year for forecasters, errors of 2 or 3 percentage points are common.

Is this record good enough? That depends on how the forecasts are used. It is certainly not good enough to support so-called *fine tuning*, that is, attempts to keep the economy always within a hair's breadth of full employment. But it probably is good enough if our interest in using discretionary stabilization policy is to close persistent and sizable gaps between actual and potential GDP.

OTHER DIMENSIONS OF THE RULES-VERSUS-DISCRETION DEBATE

While lags and forecasting play major roles in the debate between advocates of rules and advocates of discretionary policy, these are not the only battlegrounds.

THE SIZE OF GOVERNMENT

One bogus argument that is nonetheless sometimes heard is that active fiscal policy must inevitably lead to a growing public sector. Since proponents of fixed

rules tend also to oppose big government, they view this as undesirable. Of course, others think that a larger public sector is just what society needs.

This argument, however, is completely beside the point because, as we pointed out in Chapter 28 (page 660): *One's opinion about the proper size of government should have nothing to do with one's view on stabilization policy.* For example, President Ronald Reagan was as conservative as they come and devoted to *shrinking* the size of the public sector. But his tax-cutting initiatives in the early 1980s constituted an extremely *activist* policy to spur the economy. Furthermore, most stabilization policy these days is *monetary* policy, which neither increases nor decreases the size of government.

UNCERTAINTIES CAUSED BY GOVERNMENT POLICY

Advocates of rules are on stronger ground when they argue that frequent changes in tax laws, government spending programs, or monetary conditions make it difficult for firms and consumers to formulate and carry out rational plans. They argue that the authorities can provide a more stable environment for the private sector by adhering to fixed rules, which are known to businesses and consumers.

No one disputes that a more stable environment is better for private planning. But supporters of discretionary policy emphasize that stability in the economy is more important than stability in the government budget (or in Federal Reserve operations). The goal of stabilization policy is to help *prevent* gyrations in the pace of economic activity by *causing* timely gyrations in the government budget (or in monetary policy). Which atmosphere is better for business, they ask, one in which fiscal and monetary rules keep things peaceful on Capitol Hill and at the Federal Reserve while recessions and inflations rack the economy, or one in which policy instruments are changed abruptly on occasion but the economy grows more steadily? They think that the answer is self-evident.

A POLITICAL BUSINESS CYCLE?

A final argument used by advocates of rules is political rather than economic in nature. Fiscal policy, they note, is decided upon by elected politicians: the president and members of Congress. When elections are on the horizon (and for members of the House of Representatives they *always* are), politicians may be as concerned with keeping their jobs as with doing what is right for the economy. This leaves fiscal policy subject to all sorts of "political manipulations," meaning that lawmakers may take inappropriate actions to attain short-run political goals. A system of purely automatic stabilization, its proponents argue, would eliminate this peril by replacing the rule of men by the rule of law.

It is certainly *possible* that politicians could deliberately *cause* economic instability to help their own reelection. And some observers of these "political business cycles" have claimed that several American presidents have taken full advantage of the opportunity. Furthermore, even if there is no insidious intent, politicians may take the wrong actions for perfectly honorable reasons. Decisions in the political arena are never clear-cut, and it certainly is easy to find examples of grievous errors in the history of U.S. fiscal policy.

So, taken as a whole, the political argument against discretionary fiscal policy seems to have a great deal of merit. But what are we to do about it? It is unrealistic to believe that fiscal decisions could or should be made by a group of objective and nonpartisan technicians. Tax and budget policies require inherently political decisions which, in a democracy, should be made by elected officials.

This fact may seem worrisome in view of the possibilities for political chicanery. But it should not bother us any more (or any less!) than similar maneuvering in other areas of policy-making. After all, the same problem besets international relations, national defense, formulation and enforcement of the law, and so on. Politicians make all these decisions for us, subject only to sporadic accountability at elections. Is there really any reason why fiscal decisions should be different?

But monetary policy is different. Because Congress was concerned that elected officials have short time horizons, and might therefore pursue monetary policies that are too inflationary, it long ago gave the unelected technocrats at the Federal Reserve System day-to-day decision-making authority over monetary policy. There is political influence over monetary policy, but it is quite indirect: The Fed must report to Congress, and the president has the power to appoint Federal Reserve governors to his liking.

CONCLUSION: WHAT SHOULD BE DONE?

So where do we come out on the question posed at the start of this chapter? On balance, is it better to pursue the best discretionary policy we can, knowing full well that perfection cannot be achieved? Or is it wiser to rely on fixed rules and automatic stabilizers?

In weighing the pros and cons, your basic view of the economy is crucial. Some economists believe that the economy, if left unmanaged, would generate a series of ups and downs that are hard to predict, but that it would correct each of them by itself in a relatively short period of time. They conclude that, because of long lags and poor forecasts, our ability to anticipate whether the economy will need stimulus or restraint by the time policy actions have their effects is quite limited. And so they advocate fixed rules.

Other economists liken the economy to a giant glacier with a great deal of inertia. This means that, if we observe an inflationary or recessionary gap today, it is likely still to be there a year or two from now because the self-correcting mechanism works so slowly. In such a world, accurate forecasting is not imperative, even if policy lags are long. If we base policy on a forecast of a 2 percent gap between actual and potential GDP a year from now, and the gap turns out to be only 1 percent, we still will have done the right thing despite the inaccurate forecast. So holders of this view of the economy tend to support discretionary policy.

There is certainly no consensus on this issue, either among economists or among politicians. After all, the question touches upon political ideology as well as economics. And liberals have always looked to government to solve social problems, while conservatives have consistently pointed out that many efforts of government fail despite the best intentions. A prudent view of the matter might be that:

The case for active discretionary policy is strong when the economy has a serious deficiency or excess of aggregate demand. However, advocates of fixed rules are right that it is unwise to try to iron out every little wiggle in the growth path of GDP.

But one thing seems certain: The rules-versus-discretion debate seems likely to go on for quite some time.

SUMMARY

1. Monetarist and Keynesian analyses are two different ways of studying the determination of aggregate demand. Neither is a complete theory of the behavior of the economy until aggregate supply is brought into the picture.
2. **Velocity** (V) is the ratio of nominal GDP to the stock of money (M). It indicates how quickly money circulates.
3. One important determinant of velocity is the rate of interest (r). At higher interest rates, people find it less attractive to hold money because money pays no or little interest. Thus, when r rises, money circulates faster, and V rises.
4. **Monetarism** is a type of analysis that focuses attention on velocity and the money supply (M). Though monetarists realize that V is not constant, they believe that it is predictable enough to make it a useful tool for policy analysis and forecasting.
5. Because it raises output and prices, and hence increases the demand for money, expansionary fiscal policy pushes interest rates higher. This is how a monetarist explains the effect of fiscal policy. Because higher r leads to higher velocity, it leads to a higher product $M \times V$ even if M is unchanged.
6. Because fiscal policy actions affect aggregate demand either directly through G or indirectly through C, the expenditure lags between fiscal actions and their effects on aggregate demand are probably fairly short. By contrast, monetary policy operates mainly on investment, I, which responds slowly to changes in interest rates.
7. However, the policy-making lag normally is much longer for fiscal policy than for monetary policy. Hence, when the two lags are combined, it is not clear which type of policy acts more quickly.
8. Because it cannot control the demand curve for money, the Federal Reserve cannot control *both* M and r. If the demand for money changes, the Fed must decide whether it wants to hold M steady, hold r steady, or adopt some compromise position.
9. Monetarists emphasize the importance of stabilizing the growth path of the money supply while Keynesians put more emphasis on keeping interest rates on target.
10. In practice, the Fed has changed its views on this issue several times. For decades, it attached primary importance to interest rates. Between 1979 and 1982, it stressed its commitment to stable growth of the money supply. But lately the focus is clearly on interest rates again.
11. When the aggregate supply curve is very flat, increases in aggregate demand will add much to the nation's real output and add little to the price level. Under those circumstances, stabilization policy works well as an antirecession device, but has little power to combat inflation.
12. When the aggregate supply curve is steep, increases in aggregate demand increase real output rather little and succeed mostly in pushing up prices. In such a case, stabilization policy can do much to fight inflation but is not a very effective way to cure unemployment.
13. The aggregate supply curve is likely to be relatively flat in the short run but relatively steep in the long run. Hence, stabilization policy affects mainly output in the short run, but mainly prices in the long run.
14. When there are long and unpredictable *lags in the operation of fiscal and monetary policy,* attempts to stabilize economic activity may actually destabilize it.
15. The U.S. economy has a number of **automatic stabilizers** which make it less vulnerable to shocks than it would otherwise be. Among these are the personal income tax and unemployment benefits.
16. Economic forecasts are made by **econometric models,** by exploiting **leading indicators,** and by judgment. Each method seems to play a role in arriving at good forecasts. But no method is foolproof, and economic forecasts are not as accurate as we would like.
17. Some economists believe that our imperfect knowledge of the channels through which stabilization policy works, and the long lags involved, make it unlikely that discretionary stabilization policy can succeed.
18. Other economists recognize these difficulties but do not believe they are quite as serious. They also place much less faith in the economy's ability to cure recessions and inflations on its own. They therefore think that *discretionary policy* is not only advisable, but essential.
19. Stabilizing the economy by fiscal policy need *not* imply a tendency toward "big government."

KEY TERMS

Quantity theory of money
Velocity
Equation of exchange
Effect of interest rate on velocity
Monetarism
Effect of fiscal policy on interest rates
Lags in stabilization policy
Controlling M versus controlling r
Automatic stabilizers
Econometric models
Leading indicators
Judgmental forecasts
Rules versus discretionary policy

QUESTIONS FOR REVIEW

1. How much money (including cash and checking account balances) do you typically have at any particular moment? Divide this into your total income over the past 12 months to obtain your own personal velocity. Are you typical of the nation as a whole?

2. Just below, you will find data on nominal gross domestic product and the money supply (M1 definition) for selected years. Compute velocity in each year. Can you see any trend?

Year	Money Supply (M1) (billions of dollars) (end of year)	Nominal GDP (billions of dollars)
1974	274	1,497
1984	552	3,902
1994	1,149	6,931

3. Use the concept of opportunity cost to explain why velocity is higher at higher interest rates.

4. How does monetarism differ from the quantity theory of money? How does it differ from Keynesian analysis?

5. Given the behavior of velocity shown in Figure 31-1, would it make more sense for the Federal Reserve to formulate targets for M1 or M2?

6. Distinguish between the expenditure lag and the policy lag in stabilization policy. Does monetary or fiscal policy have the shorter expenditure lag? What about the policy lag?

7. Explain why their contrasting views on the shape of the aggregate supply curve lead some economists to argue much more strongly for stabilization policies to fight unemployment and others to argue much more strongly for stabilization policies to fight inflation.

8. Use a supply and demand diagram similar to Figure 31-2 (page 728) to show the choices open to the Fed following an unexpected decline in the demand for money. If the Fed is following a monetarist policy rule, what will happen to the rate of interest?

9. Explain why lags make it possible for policy actions intended to stabilize the economy actually to destabilize it.

10. Which of the following events would strengthen the argument for the use of discretionary policy, and which would strengthen the argument for rules?
    a. Structural changes make the economy's self-correcting mechanism work more quickly and reliably than before.
    b. New statistical methods are found that improve the accuracy of economic forecasts.
    c. A Democratic president is elected when there is an overwhelmingly Republican Congress. Congress and the president differ sharply on what should be done about the national economy.

11. (More difficult) Use the following hypothetical econometric model of the U.S. economy to obtain a forecast of the GDP:

$$C = 300 + 0.75DI$$
$$DI = Y - T$$
$$Y = C + I + G + (X - IM)$$

T, I, G, and $(X - IM)$ are the exogenous variables, and their forecasted values are $T = 1{,}200$, $I = 900$, $G = 1{,}300$, $(X - M) = -100$.

12. (More difficult) Answer the same question for an economy described by:

$$C = 48 + 0.9DI$$
$$DI = Y - T$$
$$T = 10 + (\tfrac{1}{3})Y$$
$$Y = C + I + G + (X - IM)$$

with forecasts $I = 200$, $G = 850$, $(X - IM) = 0$.

13. Many observers think that from 1993 to 1995 the Federal Reserve succeeded in using deft applications of monetary policy to "fine tune" the U.S. economy into the full-employment zone without worsening inflation. Use the data on money supply, interest rates, real GDP, unemployment, and the price level given on the inside back cover of this book to evaluate this claim. Ask your instructor what has happened since.

CHAPTER 32

BUDGET DEFICITS AND THE NATIONAL DEBT: FACT AND FICTION

Blessed are the young, for they shall inherit the national debt.
Herbert Hoover

There is a belief that runs deep in the American character that there is something inherently wrong with government budget deficits. Opinion polls consistently show that the public wants smaller deficits. Yet for more than a decade our political process failed to produce them. Then, after the 1992 election, things began to change. President Clinton made deficit reduction the centerpiece of his economic policy, pushing a large, multiyear package of tax hikes and expenditure cuts through a reluctant Congress in 1993. But that, apparently, just whetted political appetites, for the Republican party rode to a smashing victory in the 1994 midterm elections partly by arguing that the president had not gone nearly far enough! We had to balance the budget. Since then, the drive for a balanced budget has been at the top of the national political agenda.

What is the economic substance in this debate? What kinds of problems do large deficits pose for the economy, both now and in the future? Should we strive to balance the budget? And if so, by what means? These are the questions for this chapter.

We begin by explaining why the principles of stabilization policy that we have been learning in this part of the book do not imply that the budget should always be balanced. (Neither, however, do they lead to the conclusion that massive deficits should be the norm.) Next we look at the facts: We discuss the size of the national debt, how it grew so large, and why some economists claim that the federal budget deficit is mismeasured.

With the facts established, we examine the alleged ill effects of deficits. We shall see that many popular arguments against deficits are based on faulty reasoning. But others are not. In particular, we devote special attention to two potentially severe costs of deficit spending: It can be inflationary, and it can "crowd out" private investment spending.

PUZZLE: ARE SMALLER DEFICITS GOOD OR BAD FOR GROWTH?

The speed at which Congress should reduce or eliminate the budget deficit has been and continues to be hotly debated. One group of economists and politicians argues that the sooner we balance the budget, the better. But another group urges moderation, warning that it can be dangerous to cut the deficit too quickly.

The average person is bound to be confused by this debate. The first group argues for speed on the grounds that a lower deficit is the way to spur economic growth, while the second group urges caution on the grounds that too much deficit reduction too soon might slow the economy down! Surely, both cannot be right. Or can they? Although it sounds paradoxical, we will see in this chapter that there are elements of truth in both claims.

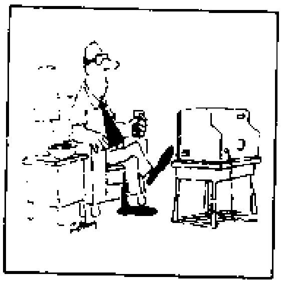
SHOULD THE BUDGET BE BALANCED?

Let us begin by reviewing the basic principles of fiscal policy that we have learned so far (especially in Chapter 28) and what they say about the goal of balancing the budget.

In brief, these principles tell us that the focus of fiscal policy should not be on *balancing the budget*, but rather on *balancing aggregate supply and aggregate demand*. They point to the desirability of budget *deficits* when private demand, $C + I + (X - IM)$, is too weak and budget *surpluses* when private demand is too strong. The budget should be balanced, according to these principles, only when $C + I + G + (X - IM)$ under a balanced-budget policy approximately equals full-employment levels of output. This may sometimes occur, but it will not necessarily be the norm.

The reason why a balanced budget is not always advisable should be clear from our earlier discussion of stabilization policy. Consider the fiscal policy that the government would follow if its goal were to balance the budget continuously. If private spending sagged for some reason, the multiplier would pull GDP down. Since personal and corporate tax receipts fall sharply when GDP declines, the budget would automatically start swinging into the red. To bring the budget back to balance, the government would then be forced either to lower spending or raise taxes—exactly the opposite of the appropriate policy response. Thus:

Attempts to balance the budget—as done, say, during the Great Depression—will prolong and deepen recessions.



"The 'Twilight Zone' will not be seen tonight, so that we may bring you the following special on the federal budget."

Budget balancing can also lead to inappropriate fiscal policy under boom conditions. If rising tax receipts induce a budget-balancing government to spend more or cut taxes, fiscal policy will "boom the boom"—with unfortunate inflationary consequences. Fortunately, believers in budget balancing usually are not alarmed by surpluses.

Actually, the issue is even more complicated than we have indicated so far. As we learned in Chapter 30, fiscal policy is not the only way the government affects aggregate demand. The government also influences aggregate demand through its monetary policy. For this reason:

The appropriate fiscal policy depends, among other things, on the stance of monetary policy. While a balanced budget may be appropriate under one monetary policy, a deficit or a surplus may be appropriate under another monetary policy.

An example will illustrate the point. Suppose Congress and the president believe that the aggregate supply and demand curves will intersect approximately at full employment if the budget is balanced. Then a balanced budget would seem to be the appropriate fiscal policy.

Now suppose monetary policy turns contractionary, pulling the aggregate demand curve inward to the left as shown in Figure 32-1 and creating a recessionary gap. If the fiscal authorities wish to restore real GDP to its original level, they must shift the aggregate demand curve back to its original position, D_0D_0. To do so, they must either raise spending or cut taxes, thereby opening up a budget deficit. Thus, the tightening of monetary policy changes the appropriate fiscal policy from a balanced budget to a deficit.

By the same token, a given target for aggregate demand implies that any change in fiscal policy will alter the appropriate monetary policy. For example, suppose Figure 32-1 indicates the effects of reducing the budget deficit by cutting government spending. Then, if the authorities do not want real GDP to fall, monetary policy must become sufficiently more expansionary to restore the aggregate demand curve to D_0D_0. Indeed, it is precisely this change in the mix of policy—a smaller budget deficit balanced by easier money—that both the Clinton administration and the Republican Congress are seeking.

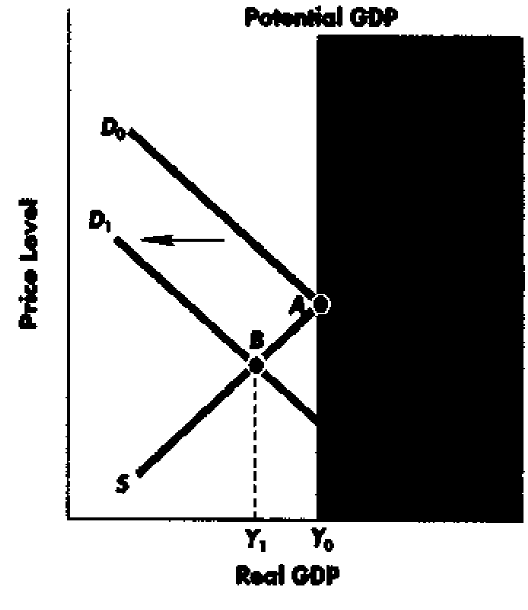
So we should not expect a balanced budget to be the norm. How, then, can we tell whether any particular deficit is too large or too small? That is a good question, but a complicated one. Before attempting an answer, we should get some facts straight.

DEFICITS AND DEBT: SOME TERMINOLOGY

First, some critical terminology. The title of this chapter contains two terms that seem similar but have different meanings: *budget deficits* and the *national debt*. We must learn to distinguish between the two.

FIGURE 32-1 THE INTERACTION OF MONETARY AND FISCAL POLICY

Both monetary and fiscal policy affect the aggregate demand curve. If monetary policy turns contractionary, the aggregate demand curve shifts inward from D_0D_0 to D_1D_1 thereby lowering real GDP from Y_0 to Y_1. If Y_0 represented full employment with a balanced budget, then Y_1 represents an economy with a recessionary gap. Expansionary fiscal policy can push the aggregate demand curve back to D_0D_0, but only by opening up a deficit in the government budget.



The **BUDGET DEFICIT** is the amount by which the government's expenditures exceed its receipts during a specified period of time, usually 1 year.

The **NATIONAL DEBT** is the federal government's total indebtedness at a moment in time. It is the result of previous deficits.

The **budget deficit** is the amount by which the government's expenditures exceed its receipts during some specified period of time, usually 1 year. For example, during fiscal year 1996, the government raised about $1,430 billion in taxes but spent almost $1,575 billion, leaving a deficit of about $145 billion.[1]

The **national debt,** also called the *public debt,* is the total value of the government's indebtedness at a moment in time. Thus, for example, the national debt at the end of fiscal year 1996 was about $5.2 trillion.

These two concepts—debt and deficits—are closely related because the government accumulates *debt* by running *deficits* or reduces its debt by running surpluses. The relationship between the debt and the deficit can be explained by a simple analogy. As you run water into a bathtub ("run a deficit"), the accumulated volume of water in the tub ("the debt") rises. Alternatively, if you let water out of the tub ("run a surplus"), the level of the water ("the debt") falls. Analogously, budget deficits raise the national debt while budget surpluses lower it. But, of course, getting rid of the deficit (shutting off the flow of water) will not get rid of the accumulated debt (drain the tub).

Having made this distinction, let us look first at the size and nature of the accumulated public debt, and then at the annual budget deficit.

SOME FACTS ABOUT THE NATIONAL DEBT

How large a public debt do we have? How did we get it? Who owns it? Is it really growing rapidly?

To begin with the simplest question, the public debt is enormous. At the end of 1996 it amounted to about $19,700 for every man, woman, and child in America. But more than one-third of this outstanding debt was owned by agencies of the U.S. government—in other words, one branch of the government owed it to another. If we deduct this portion, the net national debt was just about $3.4 trillion, or around $12,800 per person.

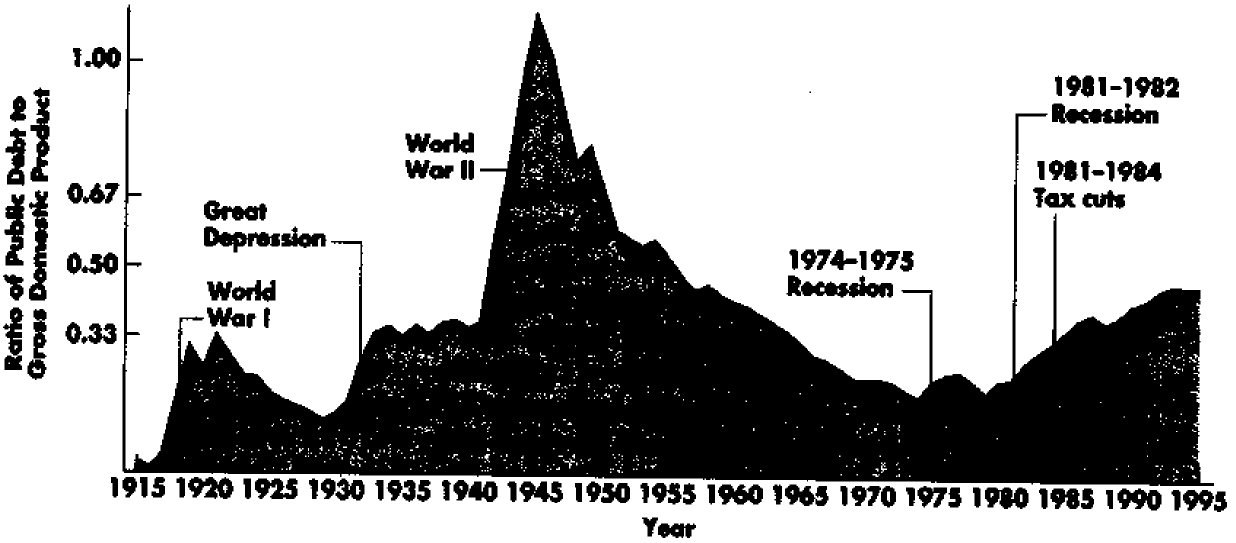
Furthermore, when we compare the debt with the gross domestic product—the volume of goods and services our economy produces in a year—it does not seem so large after all. With a GDP more than $7.5 trillion in late 1996, the net debt was under one-half of the nation's yearly income. By contrast, many families who own homes owe *several years'* worth of income to the bank that granted them mortgages. Many U.S. corporations also owe their bondholders much more than one-half of a year's sales.

But before these analogies make you feel too comfortable, we should point out that simple analogies between public and private debt are almost always misleading. A family with a large mortgage debt also owns a home with a value that presumably exceeds the mortgage. A solvent business firm has assets (factories, machinery, inventories, and so forth) that far exceed its outstanding bonds in value.

Is the same thing true of the U.S. government? No one knows for sure. How much is the White House worth? Or the national parks? And what about military bases, both here and abroad? Simply because these government assets are *not* sold on markets, no one really knows their value. However, recent federal government estimates concluded that its assets were worth "only" about $2.3 trillion at the end of 1994—a number far less than the national debt.

[1] *Reminder:* The fiscal year of the U.S. government ends on September 30. Thus, fiscal year 1996 ran from October 1, 1995, to September 30, 1996.

FIGURE 32-2 THE U.S. NATIONAL DEBT RELATIVE TO GDP, 1915–1995



This graph charts the behavior of the public debt in the United States, after (a) subtracting out the portion of the debt that is held by government agencies and (b) dividing by nominal GDP. We can see that the debt grew relative to GDP during the two world wars, during the Great Depression, during the 1974–1975 recession, and from 1981 to the early 1990s. Other than that, the debt generally has fallen relative to GDP.

SOURCE: Constructed by the authors from data in *Historical Statistics of the United States* and *Economic Report of the President*.

Figure 32-2 charts the path of the national debt from 1915 to 1995, expressing each year's national debt as a fraction of that year's nominal GDP. Looking at the debt relative to GDP is important for two reasons. First, we must remember that everything grows in a growing economy. Private debt and business debt have grown rapidly since 1915; it would have been surprising indeed if the public debt did not grow while GDP expanded so enormously. In fact, federal debt grew more slowly than private debt for most of the period from World War II until about 1980.

Second, the debt is measured in dollars and, as long as there is any inflation, the amount of purchasing power that each dollar represents declines each year. Dividing the debt by nominal GDP, as done in Figure 32-2, adjusts for both real growth and inflation, and so puts the debt numbers in better perspective.

The diagram shows us how and when the U.S. government acquired all this debt. Notice the sharp increases in the ratio of debt to GDP during World War I, the Great Depression, and especially World War II. Thereafter, you see an unmistakable downward trend until the recession of 1974–1975. In 1945, the national debt was the equivalent of 13 months' national income. By 1974, this figure had been whittled down to just 2 months' worth.

Thus, until the 1980s, most of the debt was acquired either during wars or recessions. As we shall see later, the *cause* of the debt is quite germane to the question of whether or not the debt is a burden. So it is important to remember that:

Until about 1983, almost all of the U.S. national debt stemmed from financing wars and from the losses of tax revenues that accompany recessions.

But then things changed. From the early 1980s until 1993, the national debt grew faster than nominal GDP, reversing the pattern that had prevailed since

1945. And this happened with no wars and only one recession. By 1995, the debt exceeded 5 months' GDP—nearly triple its value in 1974. This disturbing development is a major reason why economists are alarmed by continued large budget deficits.

INTERPRETING THE BUDGET DEFICIT

We have observed that the federal government's annual budget deficits have been extremely large since the Reagan administration. As Figure 32-3 shows, the budget deficit ballooned from $40 billion to the $80 billion range in fiscal years 1979 through 1981 and to $208 billion by fiscal year 1983—setting a record which was subsequently eclipsed several times. The budget for fiscal year 1996, which ended just after this edition went to press, is expected to show a deficit of $145 billion. These are enormous, even mind-boggling, numbers. But what do they mean? How are they to be interpreted?

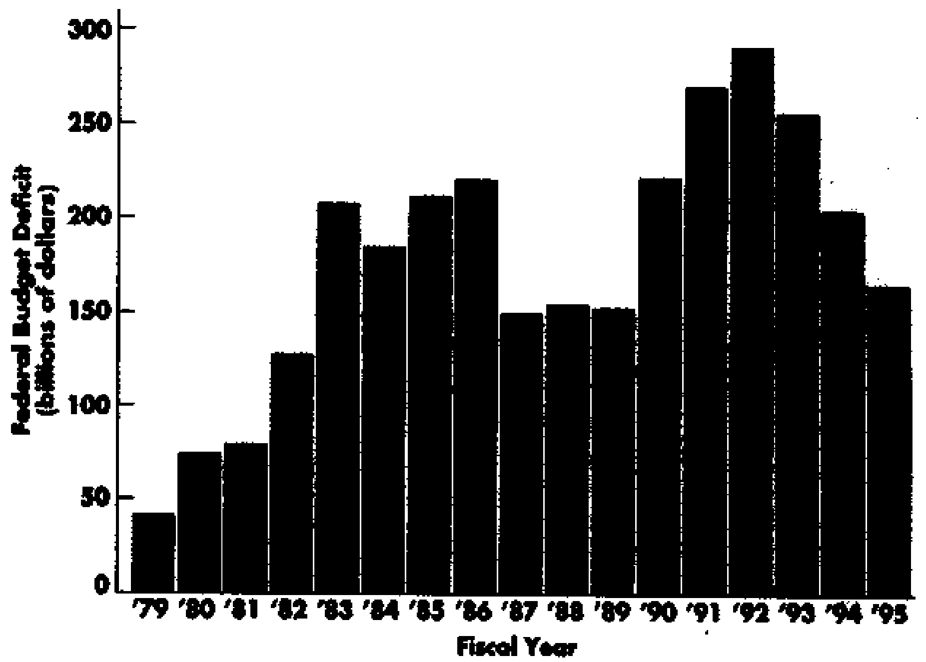
THE STRUCTURAL DEFICIT

First, it is important to understand that the same fiscal program can lead to a large or small deficit, depending on the state of the economy. Failure to appreciate this point has led many people to assume that a larger deficit always signifies a more expansionary fiscal policy. But that is not always true.

Think, for example, about what happens to the budget during a recession. As GDP falls, the government's most important sources of tax revenue—income taxes, corporate taxes, and payroll taxes—all shrink because firms and people pay lower taxes when they earn less. Similarly, some types of government spending, notably transfer payments like unemployment benefits, rise when GDP falls because more people are out of work.

FIGURE 32-3

OFFICIAL BUDGET DEFICITS, 1979–1995 (FISCAL YEARS)



The budget deficit almost tripled between fiscal year 1981 and fiscal year 1983 and did not change much between then and fiscal 1986. After dipping for a few years in the late 1980s, it then soared to a peak of $290 billion in fiscal 1992. Since then it has declined.

SOURCE: *Economic Report of the President.*

Remember that the deficit is the difference between government expenditures, which are either purchases or transfer payments, and tax receipts:

$$\text{Deficit} = G + \text{Transfers} - \text{Taxes}$$

Since a falling GDP means higher expenditures and lower tax receipts:

The deficit rises in a recession and falls in a boom, even with no change in fiscal policy.

Figure 32-4 depicts the relationship between GDP and the budget deficit. The government's fiscal program is summarized by the blue and orange lines. The horizontal blue line labeled *G* indicates that federal purchases of goods and services are approximately unaffected by GDP. The rising orange line labeled "Taxes minus Transfers" indicates that taxes rise and transfer payments fall as GDP rises. Notice that the same fiscal policy (that is, the same two lines) can lead to a large deficit if GDP is Y_1, a small deficit if GDP is Y_2, a balanced budget if GDP is Y_3, or even a surplus if GDP is as high as Y_4. Clearly, the deficit itself is not a good measure of the government's fiscal policy.

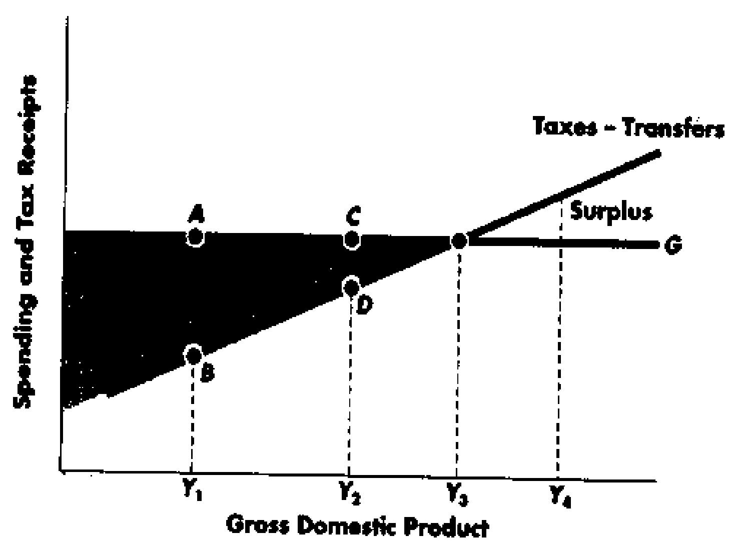
The **STRUCTURAL BUDGET DEFICIT** is the hypothetical deficit we *would have* under current fiscal policies if the economy were operating near full employment.

For this reason, many economists pay less attention to the *actual* deficit or surplus and more attention to what is called the **structural budget deficit or surplus.** This is a hypothetical measure that replaces both the spending and taxes in the *actual* budget by estimates of how much the government *would be* spending and receiving, given current tax rates and expenditure rules, if the economy were operating at some fixed, high-employment level. For example, if the high-employment benchmark in Figure 32-4 was Y_2, while actual GDP was only Y_1, the actual deficit would be *AB* while the structural deficit would be only *CD*.

Because it is based on the spending and taxing the government would be doing at some fixed level of GDP, rather than on actual expenditures and receipts, the structural deficit is insensitive to the state of the economy. It changes only when policy changes. That is why most economists view it as a better measure of the thrust of fiscal policy than the actual deficit.

This new concept helps us understand the changing nature of the large budget deficits of the 1980s. Table 32-1 displays both the actual deficit and the

FIGURE 32-4 THE EFFECT OF THE ECONOMY ON THE BUDGET



Since government purchases (*G*) do not depend on GDP, but taxes and transfer payments do, the deficit shrinks as GDP rises—even for a fixed fiscal policy. In this figure, the deficit is *AB* if GDP is Y_1, but only *CD* if GDP is Y_2. At still higher levels of GDP, the same policies could produce a balanced budget (at Y_3) or even a surplus (at Y_4).

TABLE 32-1 UNEMPLOYMENT AND THE DEFICIT

Fiscal Year	Official Deficit (billions of dollars)	+	Adjustment to High Employment (billions of dollars)	=	Structural Deficit (billions of dollars)
1981	−79		+25		−54
1983	−208		+83		−125
1985	−212		+20		−192
1987	−150		+21		−129
1989	−153		+4		−149
1991	−269		+79		−190
1993	−255		+33		−222
1995	−164		−28		−192

SOURCE Congressional Budget Office.

structural deficit for every other year since 1981. Because of recessions in 1983 and 1991, the difference between the two budgets was particularly large in those years. But it was negligible in 1989, when the economy was at full employment.

Two interesting facts stand out when we compare the numbers in the first and last columns. First, even though the official deficit was smaller in fiscal 1995 than in fiscal 1983, the structural deficit was actually larger in 1995—despite years of budget "stringency." Second, the structural deficit rose steadily from 1987 to 1993. It was this trend toward larger structural deficits that most alarmed keen students of the federal budget.

INFLATION ACCOUNTING FOR INTEREST PAYMENTS[2]

The next major problem is one of measurement rather than interpretation. Government accountants treat every dollar of interest payments as part of the government's spending. At first blush, that seems like the right thing to do, but it ignores the fundamental distinction between real and nominal interest rates that we emphasized in Chapter 23. To review the analysis:

The real interest rate tells us the amount of purchasing power the borrower turns over to the lender for the privilege of borrowing. To this we must add an *inflation premium*, equal to the expected rate of inflation, to get the *nominal interest rate*. The inflation premium compensates the lender for the expected erosion of the purchasing power of her money and is best thought of as repayment of principal.[3]

The last sentence has important implications for the government budget—implications that few people understand.

From an economic point of view, the portion of the government's interest payments that merely compensates lenders for inflation should be counted as early *repayment of principal*, not as *interest*, because it simply returns to lenders the

[2]This section contains difficult material which may be skipped in shorter courses.

[3]If you need review, see pages 550–551.

INFLATION ACCOUNTING means adjusting standard accounting procedures for the fact that inflation lowers the purchasing power of money.

purchasing power of their original funds. Only the *real* interest that the government pays should be treated as an expenditure item in the budget. Breaking up interest payments in this way is called **inflation accounting.** Since so few people understand inflation accounting, it is worth taking the time to illustrate the idea with an analogy and a simple example.

Imagine that you lend a roommate, who is enrolled in a chemistry course, a bar of radium that you happen to own. Your roommate uses the radioactive bar (carefully!) in experiments for a year, and then returns it to you. Has your loan been repaid in full? Certainly not. Because of radioactive decay, what you get back is smaller than what you originally loaned. To pay you back in full, your roommate would have to give you enough additional radium to replace the portion that eroded during the year.

The analogy to interest rates on loans is straightforward: Inflation erodes the purchasing power of money just as radioactive decay erodes radium. So, in figuring out how many dollars constitute repayment of principal, we must take inflation into account. Let's illustrate this by a concrete example: comparing a loan made at zero inflation (no decay) with a loan made at 10 percent inflation (rapid decay).

First, suppose the government borrows $1,000 for a year when the inflation rate is zero, paying 2 percent interest. At the end of the year it must pay back $1,000 in principal and $20 in interest, for a total of $1,020. Of this, only $20—the interest payment—counts as expenditure in the budget. The repayment of principal is not counted as spending. The loan transaction is summarized simply in Column 1 of Table 32-2.

Now let us see how inflation (radioactive decay of money) complicates the accountant's job. Suppose the same transaction takes place when the rate of inflation is 10 percent. To compensate the lender for 10 percent inflation, the government must return $1.10 for each dollar originally borrowed. If the real interest rate is still 2 percent, the government must return 2 percent more than this, or 1.02 × $1.10 = $1.122 per dollar borrowed. Thus, each dollar of lending earns 12.2 cents in interest, meaning that the nominal interest rate is 12.2 percent.

So suppose the government borrows $1,000 at the start of the year and repays $1,122 at year's end. Conventional accounting procedures will treat $1,000 as

TABLE 32-2

ACCOUNTING FOR A $1,000 LOAN AT A 2 PERCENT REAL INTEREST RATE

Item	(1) At Zero Inflation	(2) At 10% Inflation: Conventional Accounting	(3) At 10% Inflation: Inflation Accounting
Interest (included in budget)	$ 20	$ 122	$ 22
plus			
Principal (excluded from budget)	1,000	1,000	1,100
equals			
Total payment	$1,020	$1,122	$1,122
Addendum:			
Purchasing power of principal repayment	$1,000	$ 909	$1,000

repayment of principal (and hence not as an expenditure) and $122 as interest (which *is* an expenditure). This conventional accounting treatment is indicated in Column 2 of Table 32-2.

But these numbers are misleading since $1,000 at the end of the year is not full repayment of principal because inflation has eroded the real value of money. In fact, it is worth just $909 in beginning-of-the-year money. The correct inflation accounting treatment recognizes that it takes $1,100 at the end of the year to buy what $1,000 bought at the beginning of the year. So $1,100 is treated as repayment of principal, leaving only $22 ($1,122 − $1,100) to be treated as interest. The correct inflation accounting for the loan is shown in Column 3 of Table 32-2.

In general, the proper economic treatment of a loan in an inflationary environment must recognize that more dollars (in our example, $1,100) must be returned to the lender in order to give back the purchasing power of the original loan ($1,000). Only the excess of the nominal interest payment ($122) over the compensation for inflation ($100) should be counted as interest. Because conventional accounting does not recognize the difference between nominal and real interest:

Inflation distorts the government budget deficit under conventional accounting procedures by exaggerating interest expenses.

Our example suggests how this error can be corrected:

To correct the deficit for inflation, we must subtract the inflation premium from the interest paid on the national debt, thereby counting only *real* interest payments.

Table 32-3 shows that making the inflation adjustment to interest payments would have reduced reported deficits by about $60 billion to $90 billion in recent years, a sizable adjustment.

OTHER MEASUREMENT ISSUES

There are other complicated issues in measuring and interpreting the federal budget deficit. We conclude this section by mentioning just two.

TABLE 32-3 INFLATION ACCOUNTING AND THE DEFICIT

Fiscal Year	Official Deficit (billions of dollars)	+	Inflation Adjustment for Interest Paid (billions of dollars)	=	Inflation-Adjusted Deficit (billions of dollars)
1981	−79		+73		−62
1983	−208		+33		−175
1985	−212		+44		−168
1987	−150		+46		−104
1989	−153		+82		−71
1991	−269		+91		−178
1993	−255		+65		−190
1995	−164		+62		−102

SOURCE: Congressional Budget Office and authors' estimates.

1. *State and local budget surpluses.* Part of the reason for the federal deficit is that the federal government gives a good deal of money (around $200 billion in recent years) to state and local governments in the form of *grants-in-aid* each year. These funds help state and local governments run small annual surpluses in most years. Thus, the *combined* deficit of all levels of government is smaller than the *federal* deficit.
2. *Capital expenditures.* Some federal spending goes to purchase capital of various sorts—government buildings, military equipment, and so on. There is nothing unusual about borrowing to purchase assets. Private businesses and individuals do it all the time. For this reason, many people have suggested that the federal government compile a separate capital budget, just as most state and local governments now do.

CONCLUSION: WHAT HAPPENED AFTER 1981?

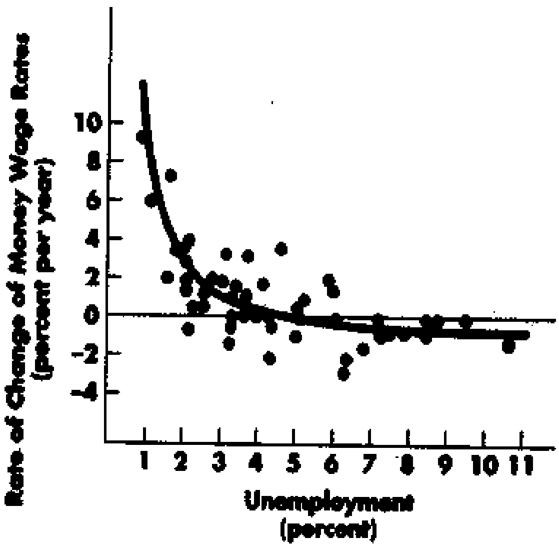
Table 32-4 puts our two major adjustments—for inflation accounting and for unemployment—together and compares the official deficits recorded since 1981 (Column 1) with the corresponding figures for the structural, inflation-corrected deficit (Column 4). The difference between the two columns is startling in some years. For example, in 1981 the economy was weak and inflation was high; the apparently substantial budget deficit of 1981 was actually a surplus on a structural, inflation-corrected basis! But since 1983 even the structural, inflation-corrected budget has been in substantial deficit.

Table 32-4 tells the following story about the recent evolution of the budget deficit. On a structural, inflation-corrected basis, the federal budget swung from a small surplus to a large deficit in the first half of the 1980s—largely due to tax

FIGURE 33-8 THE ORIGINAL PHILLIPS CURVE

This scatter diagram, reproduced from the original article by A. W. Phillips, shows the rate of change of money wages and the rate of unemployment in Great Britain between 1861 and 1913. Each year is represented by a point in the diagram.

SOURCE: A. W. Phillips, "The Relation between Unemployment and the Rate of Change of Money Wages in the United Kingdom, 1861–1957," *Economica*, New Series, vol. 25 (November 1958).



During the 1960s and early 1970s, economists often thought of the Phillips curve as a "menu" of choices available to policymakers. In this view, policymakers could opt for low unemployment and high inflation—as in 1969. Or they might prefer higher unemployment coupled with lower inflation—as, for example, in 1961. The Phillips curve, it was thought, described the *quantitative* trade-off between inflation and unemployment. And, for a number of years, it worked rather well.

Then something happened. The economy in the 1970s and early 1980s behaved far worse than the Phillips curve shown in Figure 33-9 led economists to expect. In particular, given the unemployment rates in each of those years, inflation was astonishingly high by historical standards. This is shown in Figure

FIGURE 33-9 A PHILLIPS CURVE FOR THE UNITED STATES

This Phillips curve relates *price* inflation (rather than wage inflation) to the unemployment rate in the United States for the years 1954 to 1969. Though it misses badly in a few instances (for example, 1958), it generally "fits" the data well.

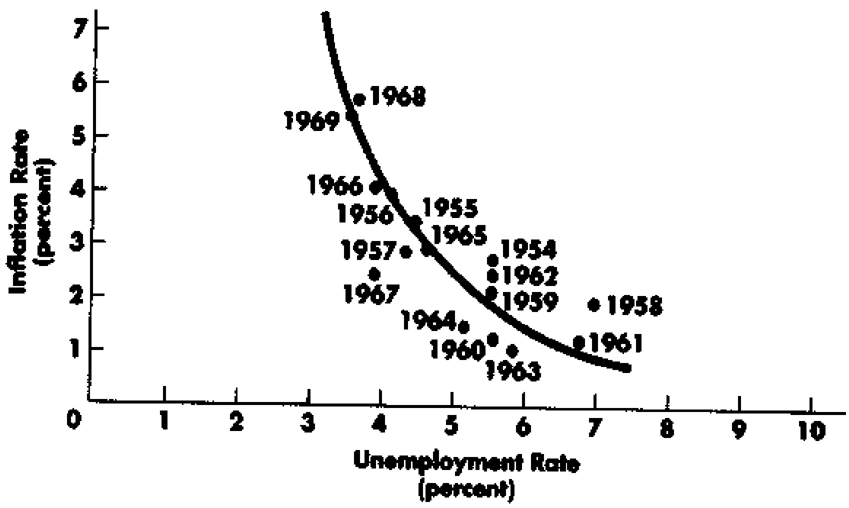
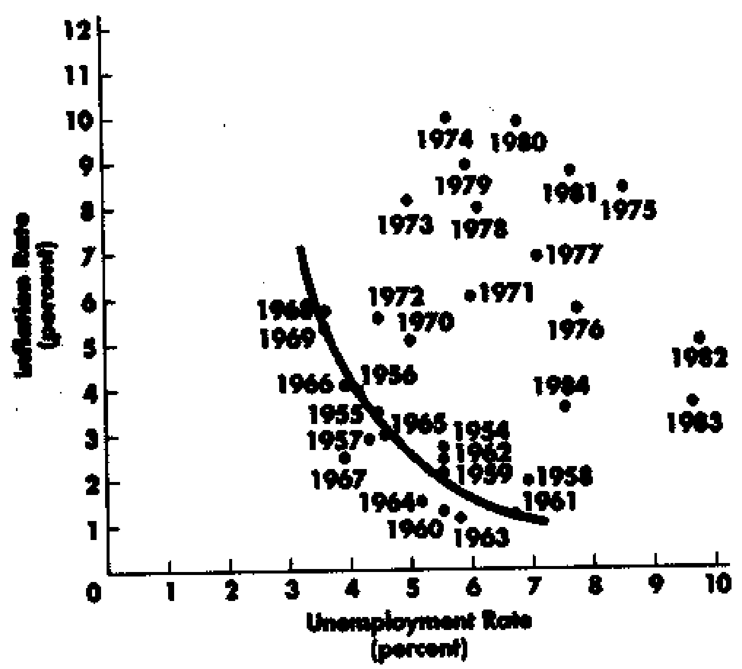


FIGURE 33-10 A PHILLIPS CURVE FOR THE UNITED STATES?

This scatter diagram adds the points for 1970 to 1984 to the scatter diagram shown in Figure 33-9. It is clear that inflation in each of those years was higher than the Phillips curve would have led us to predict.



33-10, which simply adds to Figure 33-9 the points for 1970 to 1984. Clearly something had gone wrong with the old view of the Phillips curve as a menu for policy choices. But what?

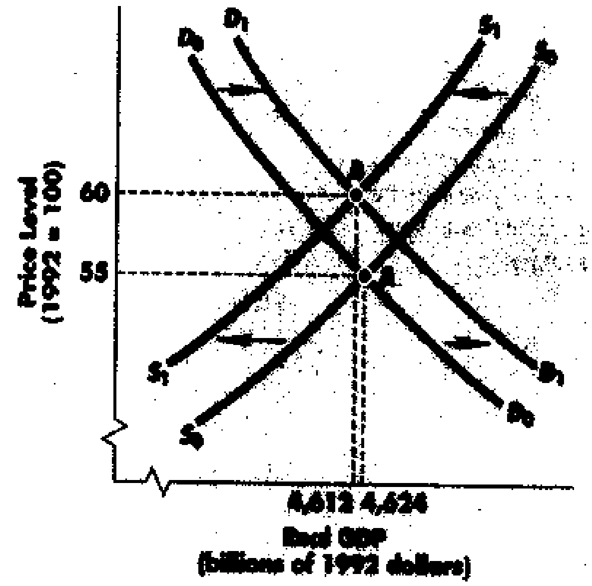
SUPPLY-SIDE INFLATION AND THE COLLAPSE OF THE PHILLIPS CURVE

There are two major answers to this question, and a full explanation contains elements of each. We begin with the simpler answer, which is that much of the inflation of the period from 1972 to 1982 did not emanate from the demand side. Instead, the 1970s and early 1980s were full of adverse "supply shocks"—events like the crop failures of 1972 to 1973 and the oil price increases of 1973 to 1974 and 1979 to 1980. These events pushed the economy's aggregate supply curve inward to the left. What kind of Phillips curve will be generated when economic fluctuations come from the supply side?

To find out, let us take the events of 1979 and 1980 as an example. In Figure 33-11, aggregate demand curve D_0D_0 and aggregate supply curve S_0S_0 represent the economic situation in 1979. Equilibrium was at point A, with a price level of 55 and real output of $4,624 billion. By 1980, the aggregate demand curve had shifted out to the position indicated by D_1D_1, and, under normal conditions, the aggregate supply curve would have shifted out as well. But 1979 to 1980 was anything but normal. The Iranian revolution led to a long shutdown of Iran's oil fields and a doubling of the price of oil.

Thus, instead of shifting *outward* as it normally does from one year to the next, the aggregate supply curve shifted *inward* from 1979 to 1980, to S_1S_1. The equilibrium for 1980 (point B in the figure) therefore wound up to the left of the equilibrium point for 1979. Real output declined slightly and prices—led by energy costs—rose rapidly.

FIGURE 33-11 STAGFLATION FROM A SUPPLY SHOCK



Instead of shifting outward as it normally does, the aggregate supply curve shifted inward—from S_0S_0 to S_1S_1—between 1979 and 1980. Coupled with fairly slow growth of the aggregate demand curve—from D_0D_0 in 1979 to D_1D_1 in 1980—equilibrium moved from point *A* to point *B*. There was a slight decline of real output, and prices rose rapidly.

Now, in a growing population with more people looking for jobs each year, a stagnant economy that is not generating new jobs suffers a rise in the unemployment rate. This is precisely what happened in the United States; the unemployment rate averaged 5.8 percent in 1979 and 7.1 percent in 1980. Thus, inflation and unemployment increased at the same time: The Phillips curve basically shifted upward. A general conclusion is that:

If fluctuations in economic activity emanate from the supply side, higher rates of inflation will be associated with higher rates of unemployment, and lower rates of inflation will be associated with lower rates of unemployment.

The major supply shocks of the 1970s stand out clearly in Figure 33-10. (Remember these are *real* data, not textbook examples.) Food prices boomed in 1972 to 1974 and again in 1978. Energy prices soared in 1973 to 1974 and again in 1979 to 1980. Clearly, the inflation and unemployment data generated by the U.S. economy in 1972 to 1974, and again in 1978 to 1980, are consistent with our model of supply-side inflation. It was supply shocks, many economists believe, that made the Phillips curve shift.

WHAT THE PHILLIPS CURVE IS NOT

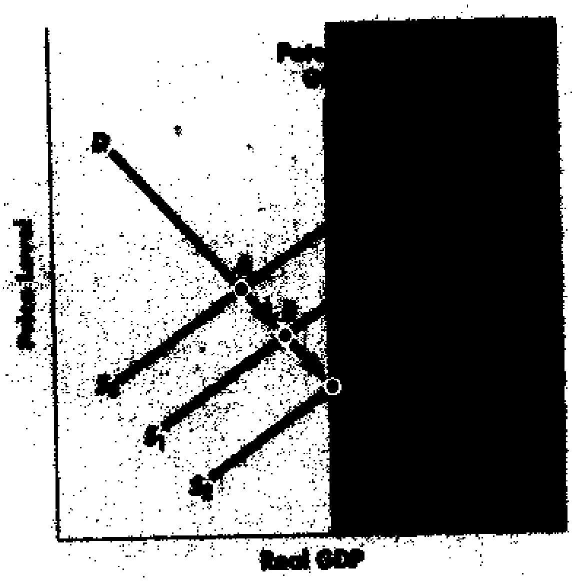
But there is another view of what went wrong in the 1970s. This one holds that policymakers misinterpreted the Phillips curve and tried to pick unsustainable combinations of inflation and unemployment.

Specifically, the Phillips curve is a *statistical relationship* between inflation and unemployment that we expect to emerge *if changes in the growth of aggregate demand are the predominant factor accounting for economic fluctuations*. But the curve was widely misinterpreted as depicting a number of *alternative equilibrium points* that the economy could achieve and from which policymakers could choose.

We can understand the flaw in this reasoning by quickly reviewing an earlier lesson. We know from Chapter 27 that the economy has a *self-correcting mecha-*

FIGURE 33-12 THE ELIMINATION OF A RECESSIONARY GAP

When the aggregate supply curve is S_0S_0 and the aggregate demand curve is DD, the economy will reach an equilibrium with a recessionary gap (point A). The resulting deflation of wages will cause the aggregate supply curve to shift outward (downward) from S_0S_0 to S_1S_1 and eventually to S_2S_2. Here, with equilibrium at point C, the recessionary gap is gone and the economy is back at normal full employment.



nism that will cure both inflations and recessions *eventually* even if the government does nothing. Why is this relevant here? Because it tells us that many combinations of output and prices cannot be maintained indefinitely. Some will "self-destruct." For example, if the economy finds itself far away from the normal full-employment level of unemployment, forces will be set in motion that tend to erode the inflationary or recessionary gap.

Figure 33-12 depicts the case of a recessionary gap where aggregate supply curve S_0S_0 intersects aggregate demand curve DD at point A. With equilibrium output well below potential GDP, there is unused industrial capacity and unsold output. So firms will not raise prices much. At the same time, the availability of unemployed workers eager for jobs limits the rate at which labor can push up wage rates. But wages are the main component of business costs, so when wages decline (relative to what they would have been without a recession) so do costs. And lower costs stimulate greater production. Figure 33-12 illustrates this idea as an outward shift of the aggregate supply curve—from S_0S_0 to S_1S_1.

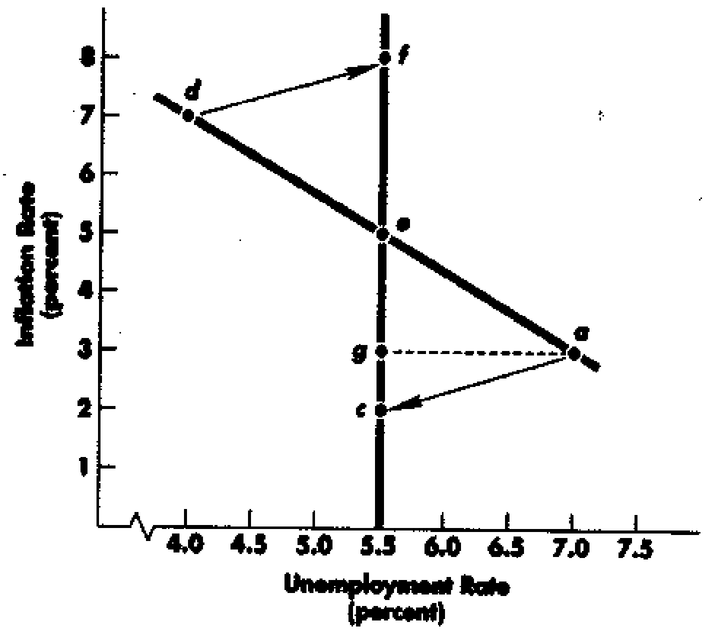
As the figure shows, the outward shift of the aggregate supply curve brought on by the recession pushes equilibrium output up as the economy moves from point A to point B. Thus, the size of the recessionary gap begins to shrink. This process continues until the aggregate supply curve reaches the position indicated by S_2S_2 in Figure 33-12. Here wages have fallen enough to eliminate the recessionary gap, and the economy has reached a full-employment equilibrium at point C.[1]

We can relate this to our discussion of the origins of the Phillips curve with the help of Figure 33-13, which is a hypothetical Phillips curve. Point a in Figure 33-13 corresponds to point A in Figure 33-12: It shows the initial recessionary gap with unemployment (assumed to be 7.0 percent) above full employment, which we assume to occur at 5.5 percent.

[1]This simple analysis assumes that the aggregate demand curve does not move during the adjustment period. If it is shifting to the right, the recessionary gap will disappear even faster, but inflation will not slow down as much. EXERCISE: Construct the diagram for this case by adding a shift in the aggregate demand curve to Figure 33-12.

FIGURE 33-13 THE VERTICAL LONG-RUN PHILLIPS CURVE

In the long run, points like *a*, where unemployment is above the normal "full-employment" unemployment rate, are unsustainable. The economy's natural self-correcting mechanism (which was described in Figure 33-12) will erode the recessionary gap by reducing both inflation and unemployment. In the diagram, this will force the economy toward a point like c. The long-run choices, therefore, are among points like *c* and *f*, which constitute what is called the vertical (long-run) Phillips curve, not among points like *d* and *a* on the downward-sloping (short-run) Phillips curve.



But we have just seen that point *A* in Figure 33-12—and therefore also point *a* in Figure 33-13—is not sustainable. The economy tends to rid itself of the recessionary gap through the disinflation process we have just described. The adjustment path from *A* to *C* depicted in Figure 33-12 would appear on our Phillips curve diagram as a movement toward less inflation and less unemployment—something like the orange arrow from point *a* to point *c* in Figure 33-13.

Similarly, points representing inflationary gaps—such as point *d* in Figure 33-13—are not sustainable. They are also gradually eliminated by the self-correcting mechanism that we studied in Chapter 27. Wages are forced up by the abnormally low unemployment, and this in turn pushes prices higher. Higher prices deter investment spending by forcing up interest rates and deter consumer spending by lowering the purchasing power of consumer wealth. The inflationary process continues until the amount people want to spend is brought into balance with the amount firms want to supply at normal full employment. During such an adjustment period, unemployment and inflation are both rising—as indicated by the orange arrow from point *d* to point *f* in Figure 33-13.

Putting these two conclusions together, we see that:

On a Phillips curve diagram, neither points corresponding to an inflationary gap (like *d* in Figure 33-13) nor points corresponding to a recessionary gap (like *a* in Figure 33-13) can be maintained indefinitely. Inflationary gaps lead to rising unemployment and rising inflation. Recessionary gaps lead to falling inflation and falling unemployment.

The economy's self-correcting mechanism always tends to push the unemployment rate back toward a specific rate of unemployment that we call the **NATURAL RATE OF UNEMPLOYMENT.**

All the points that are sustainable in the long run (such as *c*, *e*, and *f* in Figure 33-13) correspond to the same rate of unemployment, which is therefore called the **natural rate of unemployment**. The natural rate corresponds to what we have so far been calling the "full-employment" unemployment rate.

The **VERTICAL (LONG-RUN) PHILLIPS CURVE** shows the menu of inflation/unemployment choices available to society in the long run. It is a vertical straight line at the natural rate of unemployment.

Thus, the Phillips curve connecting points *d*, *e*, and *a* is not a menu of policy choices. While we can move from a point like *e* to a point like *d* by stimulating aggregate demand sufficiently, the economy will not stay at point *d*. Unemployment cannot be kept this low indefinitely. Instead, policymakers must choose from among points like *c*, *e*, and *f*, all of which correspond to the same "natural" rate of unemployment. For obvious reasons, the line connecting these points has been dubbed the **vertical (long-run) Phillips curve.** It is this vertical Phillips curve, connecting points like *e* and *f*, that represents the true long-run menu of policy choices. We thus conclude:

THE TRADE-OFF BETWEEN INFLATION AND UNEMPLOYMENT

In the short run, it is possible to "ride up the Phillips curve" toward lower levels of unemployment by stimulating aggregate demand. (See, for example, point *d* in Figure 33-13.) Conversely, by restricting the growth of demand, it is possible to "ride down the Phillips curve" toward lower rates of inflation (like point *a* in Figure 33-13). Thus, there is a *trade-off between unemployment and inflation* in the short run. Stimulating demand will improve the unemployment picture but worsen inflation; restricting demand will lower inflation but aggravate the unemployment problem.

However, *there is no such trade-off in the long run.* The economy's self-correcting mechanism ensures that unemployment eventually returns to the "natural rate," no matter what happens to aggregate demand. In the long run, faster growth of demand leads only to higher inflation, not to lower unemployment; and slower growth of demand leads only to lower inflation, not to higher unemployment.

FIGHTING UNEMPLOYMENT WITH FISCAL AND MONETARY POLICY

Now let us apply this analysis to a concrete policy problem, one that has troubled Presidents Reagan, Bush, and Clinton, and many others before them. How, if at all, should the government's ability to manage aggregate demand through fiscal and monetary policy be used to combat unemployment?

Since unemployment in the United States has been low since late 1994, we must go back to early 1993 for a real-world example. When President Clinton took office in January 1993, the inflation rate was about 3 percent and the unemployment rate was about 7 percent—a starting point just like point *a* in Figure 33-13. The president considered 7 percent unemployment intolerably high. What were his options? Should he have adopted a policy of boosting the growth of aggregate demand by expansionary fiscal and/or monetary policies?

Suppose first that nothing was done. The economy's self-correcting mechanism would have gradually eroded the recessionary gap that existed at point *a*. Both unemployment and inflation would have declined gradually as the economy moved along the orange arrow from point *a* to point *c* in Figure 33-13. Eventually, the diagram shows, the economy would have returned to its natural rate of unemployment (5.5 percent) and inflation would have fallen—from 3 percent to 2 percent in the diagram.

This eventual outcome is quite satisfactory—lower unemployment *and* lower inflation. But it may take an agonizingly long time to get there. Suppose now that the president was impatient and wanted to see unemployment decline much faster. A large dose of expansionary fiscal and monetary policy could have

pushed the economy up the short-run Phillips curve from point *a* toward point *e* in Figure 33-13. The faster economic growth and lower unemployment would have made the president (and the voters) happy. But it would also have left us with higher inflation, roughly 5 percent in the example.

This, then, is the range of choices: Wait patiently while the economy's self-correcting mechanism pulls unemployment down to the natural rate—leading to a long-run equilibrium like point *c* in Figure 33-13. Or rush the process along with expansionary stabilization policy—and wind up with the same unemployment rate but higher inflation. In what sense, then, do policymakers face a *trade-off* between inflation and unemployment? The answer is:

The cost of reducing unemployment more rapidly by expansionary fiscal and monetary policies is a permanently higher inflation rate.

WHAT SHOULD BE DONE?

Should the government pay the inflationary costs of fighting unemployment? When the transitory benefit (lower unemployment for a while) is balanced against the permanent cost (higher inflation), have we made a good bargain?

The United States opted for an intermediate strategy in 1993 to 1994. The Federal Reserve had embarked on an expansionary monetary policy to reduce unemployment well before President Clinton was elected, and it maintained this policy into early 1994. So two forces were acting simultaneously: The self-correcting mechanism was pulling the economy toward point *c* in Figure 33-13, and the Fed's expansionary policy was pushing it toward point *e*. The net result was an intermediate path (shown as the dotted line) leading to a point like *g*. By the end of 1994, the unemployment rate was down to about 5.5 percent and inflation remained around 3 percent.

How do policymakers make such decisions? Our analysis highlights three critical issues on which the answer depends.

THE COSTS OF INFLATION AND UNEMPLOYMENT

We spent an entire chapter (Chapter 23) examining the social costs of inflation and unemployment. Most of the benefits of lower unemployment, we concluded, are easily translated into dollars and cents. Basically, we need only estimate the real GDP that is gained each year. However, the costs of the permanently higher inflation rate are harder to measure. Thus, there is considerable controversy over the costs and benefits of using demand management to fight unemployment.

Some economists and public figures believe that inflation is extremely costly. They may therefore deem it unwise to accept the inflationary consequences of reducing unemployment faster. As just noted, U.S. policymakers apparently disagreed with that view in 1993.

THE SLOPE OF THE (SHORT-RUN) PHILLIPS CURVE

The shape of the short-run Phillips curve is also critical. Look back at Figure 33-13, and now suppose that the Phillips curve connecting points *a* and *e* was much steeper than shown. Then the inflationary costs of using expansionary policy to reduce unemployment would be more substantial. On the other hand, if the short-run Phillips curve was much flatter than shown in Figure 33-13, unemployment could be reduced with little inflationary cost.

THE EFFICIENCY OF THE ECONOMY'S SELF-CORRECTING MECHANISM

We have emphasized that, once a recessionary gap has opened up, it is the economy's natural self-correcting mechanism that closes it—if there is no policy response. The obvious question here is: How long do we have to wait? If the self-correcting mechanism—which works through reductions in the rate of wage inflation—is slow and halting, high unemployment will last a long time. So the costs of waiting will be enormous. But if wage inflation responds promptly, the costs of waiting will be small.

This is another issue that is surrounded by controversy. Most economists believe that the weight of the evidence points to extremely sluggish wage behavior. The rate of wage inflation appears to respond only slowly to economic slack. In terms of our Figure 33-13, this means that the economy will traverse the path from *a* to *c* at an agonizingly slow pace, so that a long period of weak economic activity will be necessary if there is to be any appreciable effect on inflation.

But a significant minority opinion finds this assessment far too pessimistic. Economists in this group argue that the costs of reducing inflation are not nearly so severe and that the key to a successful anti-inflation policy is its effects on people's *expectations*. To understand this argument, we must first examine why expectations are relevant to the Phillips-curve trade-off.

INFLATIONARY EXPECTATIONS AND THE PHILLIPS CURVE

Recall from Chapter 27 that the main reason why the economy's aggregate supply curve slopes upward—that is, why output increases as the price level rises—is that businesses typically purchase labor and other inputs under long-term contracts that fix the cost of the input in *money* terms. (The money wage rate is the clearest example.) If such contracts are in force when prices of goods go up, then *real* wages fall. Labor therefore becomes cheaper in real terms, which persuades businesses to expand employment and output. Buying cheaply and selling dearly is, after all, the route to higher profits.

Table 33-1 illustrates how this works in a concrete example. We suppose that workers and firms agree today that the money wage to be paid a year from now will be $10 per hour. The table then shows the real wage corresponding to each

TABLE 33-1 MONEY AND REAL WAGES UNDER UNEXPECTED INFLATION

Inflation Rate (percent)	Price Level 1 Year from Now	Money Wage 1 Year from Now (dollars per hour)	Real Wage 1 Year from Now (dollars per hour)
0	100	10.00	10.00
2	102	10.00	9.80
4	104	10.00	9.62
6	106	10.00	9.43

Note: Each real-wage figure is obtained by dividing the $10 nominal wage by the corresponding price level a year later and multiplying by 100. Thus, for example, when the inflation rate is 4 percent, the real wage at the end of the year is ($10.00/104) × 100 = $9.62.

alternative rate of inflation. Clearly, the higher the inflation rate, the higher the price level at the end of the year and the lower the real wage.

Lower real wages provide an incentive for the firm to increase output, as we have just noted. But lower real wages also impose losses of purchasing power on workers. Thus, there is a sense in which workers are "cheated" by inflation if they sign a contract specifying a fixed money wage in an inflationary environment.

Many economists doubt that workers will sign such contracts *if they can see inflation coming*. Would it not be wiser, these economists ask, to insist on being compensated for inflation in advance? After all, firms should be willing to offer higher money wages if they expect inflation, because they realize that higher money wages need not imply higher *real* wages. Table 33-2 illustrates how this can be done. For example, if 4 percent inflation is expected, the contract could stipulate that the wage rate be increased to $10.40 (which is 4 percent more than $10) at the end of the year. That would keep the real wage at $10, the same as it would be under zero inflation. The other money wage figures in Table 33-2 are derived similarly.

If workers and firms behave this way, and if they forecast inflation accurately, then the real wage will not decline as the price level rises. Instead, prices and wages will go up together, leaving the real wage unchanged. Workers will not lose from inflation, and firms will not gain. (In the table, the expected future real wage is $10 per hour regardless of the expected rate of inflation.) But then there would be no reason for firms to raise production when the price level rises. In a word, the aggregate supply curve would become *vertical*. In general:

If workers can see inflation coming, and if they receive compensation for it, inflation does not erode *real* wages. The economy's aggregate supply curve will then not slope upward. It will, instead, be a vertical line at the level of output corresponding to potential GDP.

Such a curve is shown in Panel (a) of Figure 33-14. Since we derived the Phillips curve from the aggregate supply curve earlier in the chapter, it follows that even the *short-run* Phillips curve would be vertical under these circumstances [see Panel (b) of Figure 33-14].[2]

If this analysis is correct, it has profound implications for the costs and benefits of inflation fighting. To see this, refer once again to Figure 33-13 on page 780, but now use the graph to depict the strategy of fighting inflation by caus-

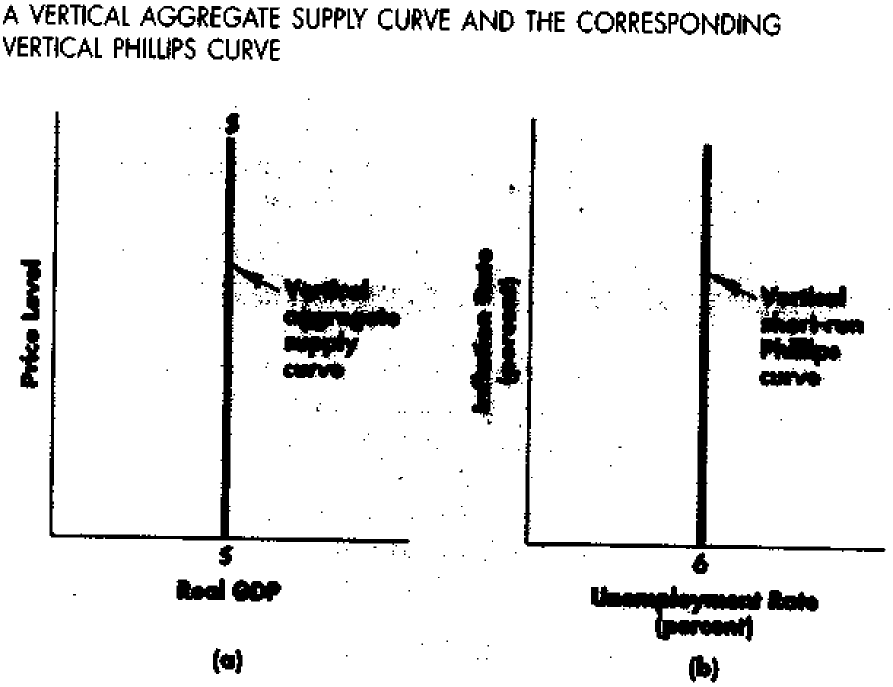
TABLE 33-2 MONEY AND REAL WAGES UNDER EXPECTED INFLATION

Expected Inflation Rate (percent)	Expected Price Level 1 Year from Now	Money Wage 1 Year from Now (dollars per hour)	Expected Real Wage 1 Year from Now (dollars per hour)
0	100	10.00	10.00
2	102	10.20	10.00
4	104	10.40	10.00
6	106	10.60	10.00

[2]See Review Question 7 at the end of the chapter.

FIGURE 33-14 A VERTICAL AGGREGATE SUPPLY CURVE AND THE CORRESPONDING VERTICAL PHILLIPS CURVE

If workers foresee inflation, and if they also receive full compensation for it in advance, then inflation will no longer erode real wages. In that case, firms will have no incentive to raise production as prices rise, and the aggregate supply curve will be vertical as in Panel (a). Since we derived the short-run Phillips curve from the aggregate supply curve, the short-run Phillips curve will also become vertical [Panel (b)].



ing a recession. Suppose we start at point *e*, with 5 percent inflation. In order to move to point *c* (representing 2 percent inflation), the economy must take a long and unpleasant detour through point *a*. Specifically, contractionary policies must push the economy down the Phillips curve toward point *a* before the self-correcting mechanism takes over and moves the economy from *a* to *c*. In other words, we must endure a recession to reduce inflation.

But what if even the *short-run* Phillips curve were *vertical* rather than downward sloping? Then this unpleasant recessionary detour would not be necessary. It would be possible for inflation to fall without unemployment rising. The economy could jump directly from point *e* to point *c*.

Is this analysis correct? Can we really slay the inflationary dragon so painlessly? Not necessarily, for our discussion of expectations so far has made at least one unrealistic assumption: that inflation can be predicted accurately. Under this assumption, as Table 33-2 shows, real wages are unaffected by inflation—leaving the aggregate supply curve vertical, even in the short run.

But forecasts of inflation are often inaccurate. Suppose workers underestimate inflation. For example, suppose they expect 4 percent inflation but actually get 6 percent. Then real wages will decline by 2 percent. More generally, real wages will fall if workers underestimate inflation *at all*. The effect of inflation on real wages will be somewhere in between those shown in Tables 33-1 and 33-2.[3] So firms will retain some incentive to raise production as the price level rises. The aggregate supply curve will remain upward sloping. We thus conclude that:

The short-run aggregate supply curve is *vertical* when inflation is predicted accurately, but *upward sloping* when inflation is underestimated. Thus, only an *unexpectedly* high inflation will raise output, because only unexpected inflation

[3]To make sure you understand why, construct a version of Table 33-2 based on the assumption that workers expect 4 percent inflation (and hence set next year's wage at $10.40 per hour), regardless of what the actual rate of inflation is. If you do this correctly, your table will show that higher inflation leads to lower real wages, as in Table 33-1.

reduces real wages. (To see this, compare Tables 33-1 and 33-2.) Similarly, only an *unexpected* decline in inflation will lead to a recession.

Since people often fail to anticipate changes in inflation correctly, this seems to leave our earlier analysis of the Phillips curve almost intact. And, indeed, most economists nowadays believe that the Phillips curve is downward sloping in the short run but vertical in the long run.

THE THEORY OF RATIONAL EXPECTATIONS

However, a vocal minority of economists—led by 1995 Nobel Prize winner Robert Lucas of the University of Chicago—disagrees. This group, believers in the hypothesis of *rational expectations*, insists that the Phillips curve is vertical even in the short run. To explain their point of view, we must first explain what rational expectations are. Then we will be in a position to understand why rational expectations have such radical implications for the trade-off between inflation and unemployment.

WHAT ARE RATIONAL EXPECTATIONS?

In many economic contexts, people must formulate expectations about what the future will bring. For example, those who invest in the stock market need to forecast the future prices of the stocks they buy and sell. And we have just discussed why workers and businesses may want to forecast future prices before agreeing on a money wage. **Rational expectations** is a controversial hypothesis about how such forecasts are made.

RATIONAL EXPECTATIONS are forecasts which, while not necessarily correct, are the best that can be made given the available data. Rational expectations, therefore, cannot err systematically. If expectations are rational, forecasting errors are pure random numbers.

As used by economists, a forecast (an "expectation") of a future variable is considered rational if the forecaster makes *optimal* use of all relevant information that is *available*. Let us elaborate on the italicized words in this definition, using as an example a hypothetical stock market investor who has rational expectations.

First, proponents of rational expectations recognize that *information is limited.* An investor interested in buying General Motors stock would like to know how much profit the company will make in the coming years. Armed with such information, she could predict the future price of GM stock more accurately. But that information is simply not available. Her forecast of the future price of GM stock is not "irrational" just because she does not know GM's future profits. On the other hand, if GM stock normally goes down on Fridays and up on Mondays, she should be aware of this fact.

Next we have the word *optimal*. As used by economists, this means using proper statistical inference to process all the relevant information that is available before making a forecast. In a word, to have rational expectations, your forecasts do not have to be correct; but they cannot have systematic errors that could be avoided by applying better statistical methods. This requirement, while exacting, is not quite as outlandish as it may seem. A good billiards player makes expert use of the laws of physics even without understanding the theory. Similarly, an experienced stock market investor may make good use of information even without formal training in statistics.

RATIONAL EXPECTATIONS AND THE TRADE-OFF

Let us now see how the hypothesis of rational expectations has been used to deny that there is any trade-off between inflation and unemployment—even in the short run.

Although they recognize that inflation cannot always be predicted accurately, rational expectationists insist that workers will not make *systematic* errors. The preceding argument about expectations tacitly assumed that workers normally *underestimate* inflation when it is rising and *overestimate* inflation when it is falling. Advocates of rational expectations claim that this is unrealistic. Workers, they argue, will always make the best possible forecast of inflation, using all the latest data and the best available economic models. Such forecasts will sometimes be too high and sometimes be too low; but they will not err systematically in one direction or the other. Consequently:

If expectations are rational, the difference between the *actual* rate of inflation and the *expected* rate of inflation (the forecasting error) must be a purely random number.

Now recall that the argument in the previous section concluded that employment is affected by inflation only to the extent that inflation *differs* from what was expected. But, under rational expectations, no *predictable* change in inflation can make the *expected* rate of inflation deviate from the *actual* rate of inflation. Hence, according to the rational expectations hypothesis, unemployment will always remain at the natural rate—except for random, and therefore totally unpredictable, gyrations due to forecasting errors. Thus:

If expectations are rational, the inflation rate can be reduced without the need for a period of high unemployment because the short-run Phillips curve is vertical.

According to the rational expectations view, the government's ability to manipulate aggregate demand gives it no ability to control real output and unemployment because the aggregate supply curve is vertical—even in the short run. [To see why, experiment by moving an aggregate demand curve when the aggregate supply curve is vertical, as in Figure 33-14(a).] Any *predictable* change in aggregate demand will lead to a change in the *expected* rate of inflation, and hence will leave real wages unaffected.

The government therefore can influence output only by making *unexpected* changes in aggregate demand. But, according to the rational expectations hypothesis, this is not easy to do because people understand what policymakers are up to. If the monetary and fiscal authorities typically react to high inflation by reducing aggregate demand, people will soon come to anticipate this reaction. And anticipated reductions in aggregate demand do not cause *unexpected* changes in inflation.

AN EVALUATION

Proponents of rational expectations are optimistic that inflation can be reduced without losses of output, even in the short run. Are they right?

As a piece of pure logic, the rational expectations argument is impeccable. Controversy arises, however, over how best to apply the idea in practice. While the theory has attracted some adherents, the evidence to date leads most economists to reject the extreme rational expectationist position in favor of the view that there is a trade-off between inflation and unemployment in the short run. There are several reasons for this.

1. Many contracts for labor and other raw materials cover such long periods of time that the expectations on which they were based, while rational at the time, may appear quite irrational from today's point of view. For example,

when 3-year labor contracts were drawn up in 1980, inflation was above 10 percent. It might therefore have been rational to expect the 1984 price level to be 30 percent above the 1981 price level and to have set 1984 wages accordingly. By late 1982, such an expectation would have been plainly irrational. But it might already have been written into contracts. If so, real wages would have wound up much higher than intended, giving firms a powerful incentive to reduce employment—even though no one behaved irrationally.

2. Many people believe that inflationary expectations do not adapt as quickly to changes in the economic environment as the rational expectations hypothesis assumes. If, for example, the government embarks on an anti-inflation policy, workers may continue to expect high inflation for quite a while. Thus, they may continue to insist on rapid money wage increases. Then, if inflation actually slows down, real wages will rise faster than anyone expected, and unemployment will result. Such behavior may not be strictly *rational*. But, to many observers, it seems realistic.
3. Some observers question whether wage agreements typically compensate workers for *expected* inflation *in advance*, as assumed by the rational expectations theory. More typically, they argue, wages catch up to *actual* inflation *after the fact*. If so, real wages will be eroded by inflation, as in the conventional view.
4. The facts have not been kind to the rational expectations point of view. The theory suggests that unemployment should hover around the natural rate most of the time, with random gyrations in one direction or the other. Yet this does not seem to be the case. The theory also predicts that preannounced (and thus expected) anti-inflation programs should be relatively painless. Yet, in practice, inflation fighting looks to be very costly. Finally, many direct tests of the rationality of expectations have cast doubt on the hypothesis. For example, survey data on people's expectations rarely meet the exacting requirements of rationality.

But all these problems with rational expectations should not obscure a basic truth. In the long run, the rational expectations viewpoint should be more or less appropriate since people will not hold incorrect expectations indefinitely. As Abraham Lincoln said, you cannot fool all of the people all of the time.

WHY ECONOMISTS (AND POLITICIANS) DISAGREE

This chapter has now taught us some of the reasons why economists often disagree about the proper conduct of national economic policy. And it also helps us understand some of the related political debates.

Should the government take stern actions to reduce inflation? You will say *yes* if you believe that (1) inflation is more costly than unemployment, (2) the short-run Phillips curve is steep, (3) expectations react quickly, and (4) the economy's self-correcting mechanism works smoothly and rapidly. These views on the economy tend to be held by monetarists and rational expectationists and by the (generally conservative) politicians who listen to them.

But you will say *no* if you believe that (1) unemployment is more costly than inflation, (2) the short-run Phillips curve is flat, (3) expectations react sluggishly, and (4) the self-correcting mechanism is slow and unreliable. These views are held by many Keynesian economists, so it is not surprising that the (generally

liberal) politicians who follow their advice often oppose the use of recession to fight inflation.

The tables turn, however, when the question is whether to use demand management to bring a recession to a rapid end. The Keynesian view of the world—that unemployment is costly, that the short-run Phillips curve is flat, that expectations adjust slowly, and that the self-correcting mechanism is unreliable—leads to the conclusion that the benefits of fighting unemployment are high while the costs are low. And so Keynesians are eager to fight recessions. The monetarist and rational expectationist positions on these four issues are precisely the reverse, and so are the policy conclusions.

THE DILEMMA OF DEMAND MANAGEMENT

So we have seen that the makers of monetary and fiscal policy face an agonizing trade-off. If they stimulate aggregate demand to reduce unemployment, they will aggravate inflation. If they restrict aggregate demand to fight inflation, they will cause higher unemployment.

But wait. Early in the chapter we learned that when inflation comes from the supply side, inflation and unemployment are *positively* associated: We suffer from more of both or enjoy less of each. Does this mean that monetary and fiscal policymakers can escape the trade-off between inflation and unemployment? Certainly not.

THE TRADE-OFF BETWEEN INFLATION AND UNEMPLOYMENT

Adverse shifts in the aggregate supply curve can cause both inflation and unemployment to rise together, and thus can destroy the statistical Phillips curve relationship. Nevertheless, anything that monetary and fiscal policy can do will make unemployment and inflation move in *opposite* directions because monetary and fiscal policy give the government control only over the *aggregate demand* curve, not over the *aggregate supply* curve. Thus, no matter what the source of inflation, and no matter what happens to the Phillips curve, the monetary and fiscal policy authorities must still face up to the disagreeable trade-off between inflation and unemployment. This is a principle that many policymakers have failed to recognize, and one of the **Ideas** that we hope you will remember well **Beyond the Final Exam.**

Naturally, the unpleasant nature of this trade-off has led both economists and public officials to search for a way out of the dilemma. The rest of this chapter—including the Policy Debate box on the next page—considers some of these ideas, none of which are panaceas.

ATTEMPTS TO REDUCE THE NATURAL RATE OF UNEMPLOYMENT

One highly desirable approach—if only we knew how to do it—would be to reduce the natural rate of unemployment. Then we could enjoy lower unemployment without higher inflation. The question is: How?

The most promising approaches have to do with education, training, and job placement. The data clearly show that more educated workers are unemployed

POLICY DEBATE — SHOULD THE FED CONCENTRATE ONLY ON REDUCING INFLATION?

One disarmingly simple way to sidestep the dilemma of demand management is to eliminate one of the two traditional goals. In particular, some economists and politicians have proposed that the Federal Reserve be directed to concentrate on price stability and forget about trying to stabilize output and employment. The argument runs something like this.

1. Price stability is of overriding importance to the smooth functioning of a market economy.
2. The short-run trade-off between inflation and unemployment impedes the Fed's efforts to reduce inflation. (In more extreme versions of the argument, this step denies the existence of any short-run trade-off.)
3. The Fed has failed in its efforts to stabilize output and employment and, in the process, it has lost ground in the war against inflation.

Let us briefly examine each step of the argument.

The first proposition is a matter of intense debate, as we have emphasized in this and earlier chapters. Someone who believes that inflation is a terrible problem and unemployment a minor one might very well want the Fed to concentrate on reducing inflation. But other people hold the opposite view.

The second proposition is undoubtedly true. Concerns about the unemployment consequences of its actions surely have tempered the Fed's inflation-fighting zeal on many occasions.

The third proposition is again debatable. As we have learned in previous chapters, there have been episodes in which the Fed's stabilization policy has been stunningly successful (such as 1993 to 1995) and episodes in which it has been a miserable failure (such as the 1930s and 1972 to 1973). Lately, however, most observers give the Fed good marks.

In 1995, legislation was introduced that would direct the Fed to make price stability its primary goal. But the bill has yet to reach the floor of either house of Congress.

less frequently than less educated ones. Vocational training and retraining programs, if successful, help unemployed workers with obsolete skills acquire abilities that are currently in demand. By so doing, they both raise employment and help alleviate upward pressures on wage rates in jobs where qualified workers are in short supply.

Government job placement services play a similar role. Such programs try to improve the match of workers to jobs by funneling information from prospective employers to prospective employees.

These ideas sound promising and sensible. But there are two big problems. First, training and placement programs often look better on paper than they do in practice, where they have achieved only modest success. Too often, people are trained for jobs that do not exist by the time they finish their training—if indeed they ever existed.

The second problem is that the high cost of these programs restricts the number of workers that can be accommodated, even when they work. For this reason, publicly supported job training is done on a very small scale in the United States—much less, say, than in most European countries. Such small expenditures can hardly be expected to make a significant dent in the natural rate of unemployment. Here job placement and counseling may have an important edge because the cost per unemployed worker is far less than the cost of retraining.

The Clinton administration has vigorously promoted initiatives in both retraining and placement; the expressed goal is to transform the unemployment system into a reemployment system. But, so far, Congress has been reluctant to go along.

WAGE-PRICE CONTROLS

WAGE-PRICE CONTROLS are legal restrictions on the ability of industry and labor to raise wages and prices.

To many people, the most natural way to control inflation is to impose mandatory **wage–price controls.** After all, if we do not like something, why not just outlaw it? Then the government could concentrate on fighting unemployment without worrying about the trade-off.

Economists generally oppose controls for reasons we learned back in Chapter 4. When price ceilings are effective, they force prices below equilibrium levels, thereby causing shortages. With prices no longer serving as rationing devices, something else must take their place. One possibility is long lines of buyers waiting their turn, as was typical in the former Soviet bloc. Another is government ration coupons, a device used successfully in the United States during World War II.

But neither of these measures is likely to be popular with the electorate in peacetime. And both are likely to spawn black markets, which erode respect for law and order at the same time they undermine the effects of controls. As critics are fond of pointing out, controls give perfectly law-abiding citizens incentives to break the law in an effort to circumvent the controls.

Wage–price controls were last used in the United States in the early 1970s under President Nixon, and have acquired a bad name. But some foreign countries like Argentina and Israel have used controls successfully to engineer dramatic reductions of inflation.

INDEXING

INDEXING refers to provisions in a law or a contract whereby monetary payments are automatically adjusted whenever a specified price index changes. Wage rates, pensions, interest payments on bonds, income taxes, and many other things can be indexed in this way, and have been. Sometimes such contractual provisions are called *escalator clauses*.

Indexing—which refers to provisions in a law or contract whereby monetary payments are automatically adjusted whenever a specific price index changes—presents a very different approach to the inflation–unemployment dilemma. Instead of trying *to lower the inflation rate*, the primary purpose of indexing is *to reduce the social costs of inflation.*

The most familiar example of indexing in the United States is an *escalator clause* in a wage agreement. An escalator clause provides for an automatic increase in money wages—without the need for new contract negotiations—any time the price level rises by more than a specified amount. In this way, workers are partly protected from inflation. Nowadays, with inflation low and stable, less than half of all workers employed by large, unionized firms, and very few nonunion workers, are covered by escalator clauses.

Interest payments on bonds or savings accounts can also be indexed, although this is not currently done in the United States.[4] The most extensive indexing to be found in the United States today is in government transfer payments. Social Security benefits, for instance, are indexed so that retirees are not victimized by inflation.

Some economists believe that the United States should follow the example of several foreign countries and adopt a much more widespread system of indexing. Why? Because, they argue, it would take most of the sting out of inflation. To see how, let us review some of the social costs of inflation that we enumerated in Chapter 23.

One important cost is the capricious redistribution of income caused by unexpected inflation. We saw that borrowers and lenders normally incorporate an *inflation premium* equal to the *expected rate of inflation* into the nominal interest

[4] Some other countries, with much higher inflation than ours, do extensive indexing of interest rates. Brazil and Israel are notable examples.

rate. Then, if inflation turns out to be higher than expected, the borrower has to pay the lender only the agreed-upon nominal interest rate, including the premium for *expected* inflation; he does not have to compensate the lender for the (higher) *actual* inflation. Thus, the borrower enjoys a windfall gain and the lender loses out. The opposite happens if inflation turns out to be lower than expected.

But if interest rates on loans were indexed, none of this would occur. Borrowers and lenders would agree on a fixed *real* rate of interest, and then the borrower would compensate the lender for whatever *actual inflation* occurred. No one would have to guess what the inflation rate would be.[5]

A second social cost we mentioned in Chapter 23 stems from the fact that our tax system levies taxes on nominal interest and nominal capital gains. As we learned, this flaw in the tax system leads to extremely high effective tax rates in an inflationary environment. But indexing could fix this problem easily. We need only rewrite the tax code so that only *real* interest payments and *real* capital gains are taxed.

A final problem noted in Chapter 23 is that uncertainty over future price levels makes it difficult to enter into long-term contracts—wage agreements, rental agreements, construction agreements, and so on. One way out of this problem is to write indexed contracts, which specify all future payments in real terms.

In the face of all these benefits, why do many economists oppose indexing? Probably the major reason is the fear that indexing will lead to an acceleration of inflation. With the costs of inflation reduced so markedly, they argue, what will persuade governments to pay the price of fighting inflation? What will stop them from inflating more and more? They fear that the answer to these questions is: Nothing. Voters who stand to lose nothing from inflation are unlikely to pressure their legislators into stopping it. Opponents of indexing worry that a mild inflationary disease could turn into a ravaging epidemic in a highly indexed economy.

[5]For example, an indexed loan with a 2 percent real interest rate would require a 5 percent nominal interest payment if inflation were 3 percent, a 7 percent nominal interest payment if inflation were 5 percent, and so on.

SUMMARY

1. Inflation can be caused either by rapid growth of aggregate demand or by sluggish growth of aggregate supply.

2. When fluctuations in economic activity emanate from the demand side, prices will rise rapidly when real output grows rapidly. Since rapid growth means more jobs, unemployment and inflation will be inversely related.

3. This inverse relationship between unemployment and inflation is called the **Phillips curve.** U.S. data for the 1950s and 1960s display a Phillips-curve relation, but data for the 1970s and 1980s do not.

4. The Phillips curve is not a menu of *long-run* policy choices for the economy because the *self-correcting mechanism* guarantees that neither an inflationary gap nor a recessionary gap can last indefinitely.

5. Because of the self-correcting mechanism, the economy's true long-run choices lie along a **vertical long-run Phillips curve,** which shows that the so-called **natural rate of unemployment** is the only unemployment rate that can persist indefinitely.

6. In the short run, the economy can move up or down its short-run Phillips curve. *Temporary* reductions in unemployment can be achieved at the cost of higher inflation. Similarly, *temporary* increases in unemployment can be used to fight inflation.

7. Whether it is advisable to use unemployment to fight inflation depends on four principal factors: the relative social costs of inflation versus unemployment, the efficiency of the economy's self-correcting mechanism, the shape of the short-run Phillips curve, and how quickly inflationary expectations adjust.

8. If workers expect inflation to occur, and if they demand (and receive) compensation for inflation, output will be independent of the price level. Both the aggregate supply curve and the short-run Phillips curve are vertical in this case.
9. However, errors in predicting inflation will still change real wages and therefore the quantity of output that firms wish to supply. Thus, *unpredicted* movements in the price level will lead to a normal, upward-sloping aggregate supply curve.
10. According to the hypothesis of **rational expectations,** errors in predicting inflation are purely random. This means that, except for some random (and uncontrollable) gyrations, the aggregate supply curve is vertical even in the short run.
11. Many economists reject the rational expectations view of the world. Some deny that expectations are "rational" and believe instead that people tend, for example, to underpredict inflation when it is rising. Others point out that contracts signed years ago cannot possibly embody expectations that are "rational" in terms of what we know today.
12. When fluctuations in economic activity are caused by shifts of the aggregate supply curve, output will grow slowly (causing unemployment to rise) when inflation speeds up. Hence, the rates of unemployment and inflation will be positively related. Many observers feel that this sort of *stagflation* is why the Phillips curve collapsed in the 1970s.
13. Even if inflation is initiated by supply-side problems, so that inflation and unemployment occur together, the monetary and fiscal authorities still face this trade-off: Anything they do to improve unemployment is likely to worsen inflation, and anything they do to reduce inflation is likely to aggravate unemployment. The reason is that monetary and fiscal policy mainly influence the aggregate demand curve, not the aggregate supply curve. This is one of our **Ideas for Beyond the Final Exam.**
14. Policies that improve the functioning of the labor market—including retraining programs and employment services—can in principle lower the natural rate of unemployment. To date, however, the U.S. government has had only modest success with these measures.
15. **Wage–price controls** reduce inflation quite directly. But legal limits on wage and price increases seriously interfere with the workings of a market economy.
16. **Indexing** is another way to approach the trade-off problem. Instead of trying to improve the trade-off, it concentrates on reducing the social costs of inflation—perhaps eliminating them altogether. Opponents of indexing worry, however, that the economy's resistance to inflation may be lowered by indexing.

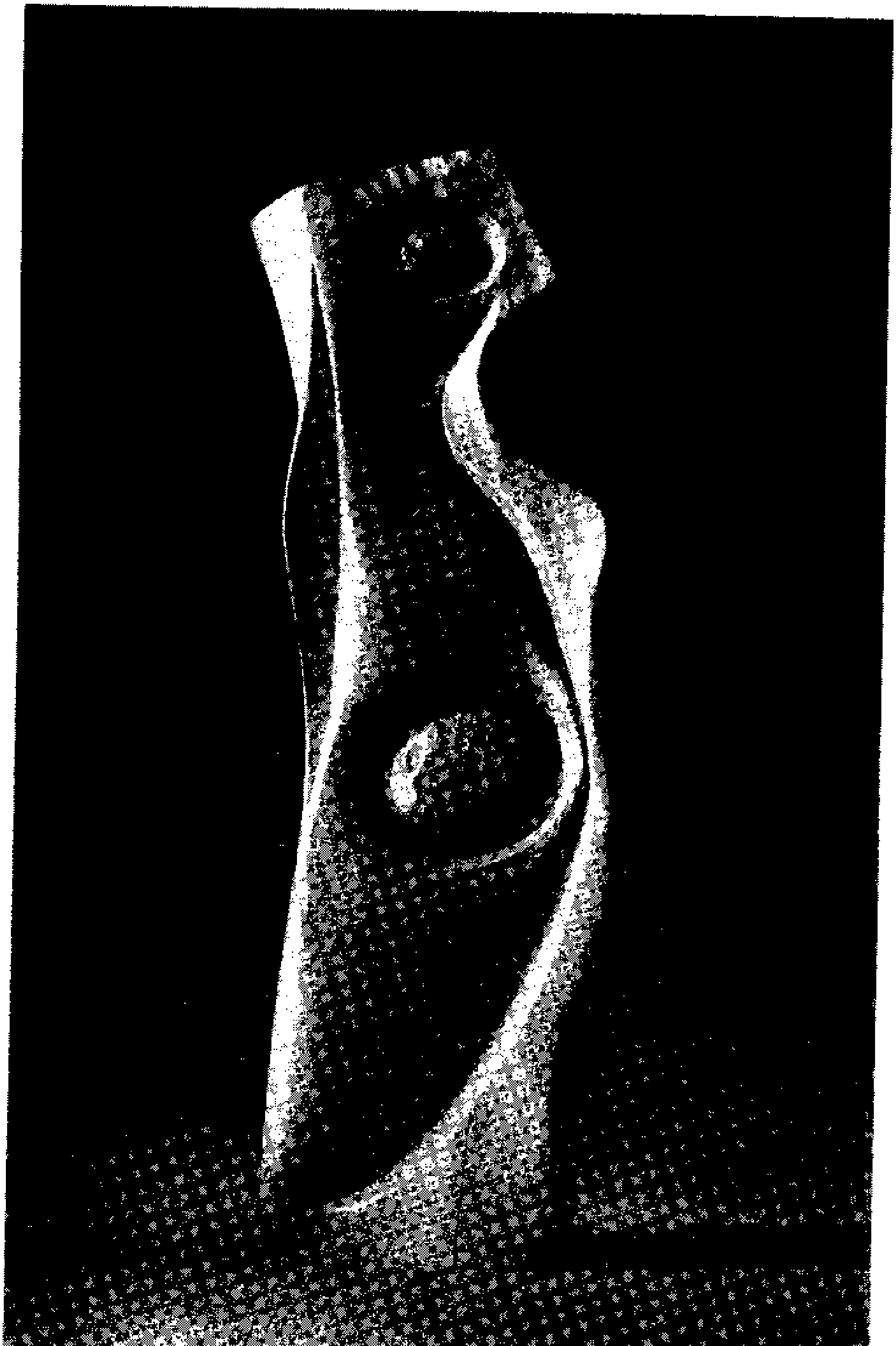
KEY TERMS

Demand-side inflation
Supply-side inflation
Phillips curve
Stagflation caused by supply shocks
Self-correcting mechanism
Natural rate of unemployment
Vertical (long-run) Phillips curve
Trade-off between unemployment and inflation in the short run and in the long run
Rational expectations
Wage–price controls
Indexing (escalator clauses)
Real versus nominal interest rates

QUESTIONS FOR REVIEW

1. Seeing that inflation and unemployment sometimes rise together, some observers deny that policymakers face a trade-off between inflation and unemployment. Are they correct?
2. "There is no sense in trying to shorten recessions through fiscal and monetary policy because the effects of these policies on the unemployment rate are sure to be temporary." Comment on both the truth of this statement and its relevance for policy formulation.
3. Why is it said that decisions on fiscal and monetary policy are, at least in part, political decisions that cannot be made on "objective" economic criteria?
4. What is a "Phillips curve"? Why did it seem to work so much better in the 1954 to 1969 period than it did in the 1970s?
5. Explain why expectations about inflation affect the wages that result from labor-management bargaining.
6. What is meant by "rational" expectations? Why does the hypothesis of rational expectations have such stunning implications for economic policy? Would believers in rational expectations want to shorten a recession by expanding aggregate demand? Would they want to fight inflation by reducing aggregate demand?

7. Show that, if the economy's aggregate supply curve is vertical, fluctuations in the growth of aggregate demand produce only fluctuations in inflation with no effect on output. Relate this to your answer to the previous question.
8. Long-term government bonds now pay approximately 6 percent *nominal* interest. Would you prefer to trade yours in for an indexed bond that paid a 3 percent *real* rate of interest? What if the real interest rate offered were 2 percent? What if it were 1 percent? What do your answers to these questions reveal about your personal attitudes toward inflation?
9. In the late 1980s, the unemployment rate in the United States hovered in the 5.0 to 5.5 percent range, and the inflation rate rose. In the mid-1990s, unemployment fluctuated between 5.5 percent and 6.0 percent, and inflation was stable. What do these facts suggest about the numerical value of the natural rate of unemployment?
10. It is said that the Federal Reserve Board typically cares more about inflation and less about unemployment than the administration. If this is true, why might President Clinton have been worried about what Fed Chairman Alan Greenspan would do in early 1995, when inflation increased for a few months?
11. The year 1996 opened with the unemployment rate under 6 percent, real GDP growing moderately, inflation about 3 percent, and the federal budget deficit around $145 billion.
    a. Make an argument for engaging in contractionary monetary or fiscal policies under these circumstances.
    b. Make an argument for engaging in expansionary monetary or fiscal policies under these circumstances.
    c. Which argument do you find more persuasive?



PART VIII

The United States in the World Economy

CHAPTER 34

INTERNATIONAL TRADE AND COMPARATIVE ADVANTAGE

No nation was ever ruined by trade.
Benjamin Franklin



International trade is vital to the health of any nation, and so it is essential to our study of economics. The world's major economies have always been linked in various ways. But dramatic improvements in transportation, telecommunications, and international relations in recent decades have drawn the industrial nations of the world ever closer together.

Economic events in other countries affect our economy for both macroeconomic and microeconomic reasons. For example, we learned in Parts VI and VII that the level of net exports is one important determinant of a nation's output and employment. But we have not delved very deeply into the factors that determine a nation's exports and imports. Chapters 35 and 36 will take up these *macroeconomic* linkages in greater detail.

But, first, this chapter studies some of the reasons why international trade is important to a nation's *microeconomic* well-being. The central principle here is the *law of comparative advantage*, which plays a major role in determining the patterns of world trade. We will also learn how the prices of internationally traded goods are determined by supply and demand in a free world market. Finally, we will examine the effects of government interferences with foreign trade through quotas, tariffs, and other devices designed to protect domestic industries from foreign competition.

PUZZLE: HOW CAN AMERICANS COMPETE WITH "CHEAP FOREIGN LABOR"?

Why do Americans (like the citizens of other nations) often want their government to limit or prevent import competition? One major reason is the common belief that imports take bread out of the mouths of American workers. According to this view, "cheap foreign labor" steals jobs from Americans and puts pressure on U.S. businesses to lower wages.

Unfortunately, the facts are not consistent with this story. For one thing, wages in industrialized countries that export to the United States have risen spectacularly during the past 25 years. Table 34-1 shows that wages in seven leading

TABLE 34-1

LABOR COSTS IN INDUSTRIALIZED COUNTRIES

	1970	1994
	(percentage of U.S. labor costs)	
France	41	100
United Kingdom	35	80
Italy	42	95
Japan	24	125
Netherlands	51	122
Sweden	70	110
(West) Germany	56	160

Data are compensation estimates per hour and relate to production workers in the manufacturing sector.

SOURCE U.S. Bureau of Labor Statistics.

countries rose from an average of only 46 percent of American wages in 1970 to 113 percent by 1994. By 1994, labor costs in Sweden, the Netherlands, western Germany, and Japan exceeded our own, and costs in France were about equal. Yet American imports of Toyotas from Japan, Volkswagens from Germany, and Volvos from Sweden grew as wages in those countries rose relative to American wages.

By comparison, when European and Japanese wages were far below those in the United States in the 1950s, American industry had no trouble marketing our products abroad. In fact, the main problem then was to bring our imports up to the level at which they roughly balanced our bountiful exports. Ironically, our position in the international marketplace deteriorated as wage levels in Europe and Japan began to rise closer to our own.

Clearly, then, cheap foreign labor must not be a major obstacle to U.S. sales abroad—as a "common-sense" view of the matter suggests. In this chapter we will see what is wrong with that view.

WHY TRADE?

The earth's resources are not equally distributed across the planet. While the United States can produce its own coal and wheat, it is almost *entirely* dependent on the rest of the world for such items as rubber and coffee. Similarly, Saudi Arabia has little land that is suitable for farming, but sits atop a huge pool of oil. Because of the seemingly whimsical distribution of vital resources, every nation must trade with others to acquire what it lacks.

Even if countries had all the resources they needed, other differences in natural endowments—such as climate, terrain, and so on—would lead them to engage in trade. Americans *could*, with great difficulty, grow their own banana trees and coffee shrubs in hothouses. But these crops are much more efficiently grown in Honduras and Brazil, where the climates are appropriate.

The skills of a nation's labor force also play a role. If New Zealand has a large group of efficient farmers and few workers with industrial experience while the opposite is true in Japan, it makes sense for New Zealand to specialize in agriculture and let Japan concentrate on manufacturing.

Finally, a small country that tried to produce every product would end up with many industries too small to utilize mass-production techniques and other methods that confer cost advantages on large-scale operations. For example, some countries operate their own international airlines for reasons that can only be political, not economic. Inevitably, small nations that insist on competing in industries that are economical only when their scale of operation is large find that these enterprises can survive only with the aid of large government subsidies.

To summarize, the main reason why nations trade with one another is to exploit the many advantages of **specialization.**

SPECIALIZATION means that a country devotes its energies and resources to only a small proportion of the world's productive activities.

International trade is essential for the prosperity of trading nations because:

1. Every country lacks some vital resources that it can get only by trading with others.
2. Each country's climate, labor force, and other endowments make it a relatively efficient producer of some goods and an inefficient producer of other goods.
3. Specialization permits larger outputs and can therefore offer economies of large-scale production.

MUTUAL GAINS FROM TRADE

Many people believe that a nation can gain from trade only at the expense of another. Centuries ago, writers on international trade pointed out that nothing new is produced by the mere act of trading. They therefore incorrectly argued that, if one country gains from a swap, the other country must necessarily lose. One of the consequences of this mistaken belief was and is a policy prescription calling for each country to try to disadvantage its trading partners—on the grounds that one nation's gain must be another's loss.

Yet, as Adam Smith emphasized, and as we learned in Chapter 3, both parties *must* expect to gain something from any *voluntary exchange*. Otherwise why would they agree to trade?

But how can mere exchange of goods leave *both* parties better off? The answer is that while trade does not increase the total output of goods, it does allow each party to acquire items better suited to its tastes. Suppose Scott has four cookies and nothing to drink, while William has two glasses of milk and nothing to eat. A trade of two of Scott's cookies for one of William's glasses of milk does not increase the total supply of either milk or cookies, but it almost certainly improves the welfare of both boys.

By exactly the same logic, both the United States and Mexico must be better off if Mexico voluntarily ships tomatoes to the United States in return for chemicals. In general:

MUTUAL GAINS FROM VOLUNTARY EXCHANGE

Both parties must expect to gain from any *voluntary exchange*. Trade brings about mutual gains by redistributing products so that both parties end up holding more preferred combinations of goods than they held before. This principle, which is one of our **Ideas for Beyond the Final Exam,** applies to nations just as it does to individuals.

INTERNATIONAL VERSUS INTRANATIONAL TRADE

The 50 states of the United States may be the most eloquent testimony to the gains from specialization and free trade. Florida specializes in growing oranges, Iowa in growing corn, Pennsylvania makes steel, and Michigan builds cars. All these states trade freely with one another and enjoy great material prosperity. Try to imagine how much lower your standard of living would be if you could consume only items produced in your own state.

The essential logic of international trade is no different from that underlying trade among different states; the basic reasons for trade are equally applicable *within* a country or *among* countries. Why, then, do we study international trade as a special subject? There are at least three reasons.

POLITICAL FACTORS IN INTERNATIONAL TRADE

First, domestic trade takes place under a single national government, while foreign trade always involves at least two governments. But a nation's government is normally much less concerned about the welfare of citizens of other countries than it is about its own citizens. So, for example, the Constitution of the United States prohibits tariffs on trade among states but does not prohibit the United States from imposing tariffs on imports from abroad. One major issue in the economic analysis of international trade is the use and misuse of impediments to free international trade.

THE MANY CURRENCIES INVOLVED IN INTERNATIONAL TRADE

Second, all trade within the borders of the United States is carried out in U.S. dollars. But trade across national borders must involve at least two currencies. Rates of exchange between different currencies can and do change. In 1985, it took about 250 Japanese yen to buy a dollar; now it takes only about 105. Variability in exchange rates brings with it a host of complications and policy problems that are discussed in Chapters 35 and 36.

IMPEDIMENTS TO MOBILITY OF LABOR AND CAPITAL

Third, it is much easier for labor and capital to move about within a country than to move from one country to another. If jobs are plentiful in Michigan but scarce in West Virginia, workers can move freely to follow the job opportunities. Of course, there are personal costs, including the financial cost of moving and the psychological cost of leaving friends and familiar surroundings. But such relocations are not inhibited by immigration quotas, by laws restricting the employment of foreigners, nor by the need to learn new languages.

There are also greater impediments to the transfer of capital across national boundaries than to its movement within a country. For example, many countries have rules limiting foreign ownership; even the United States limits foreign ownership of broadcast outlets and airlines. Foreign investment is also subject to special political risks, such as the danger of outright expropriation after a change in government. For example, in 1994 U.S. citizens had more than 1,300 unresolved expropriation claims against the government of Nicaragua.

But even if nothing as extreme as expropriation occurs, capital invested abroad faces significant risks from changes in exchange rates. An investment valued at 250 million yen will be worth $1.0 million to American investors if the dollar is worth 250 yen, but $2.5 million if the dollar is worth just 100 yen.

THE LAW OF COMPARATIVE ADVANTAGE

The gains from international specialization and trade are clear when one country is better at producing one item while its trading partner is better at producing another. For example, no one finds it surprising that Brazil sells coffee to the United States while America exports aircraft to Brazil. We know that coffee can be produced using less labor and other inputs in Brazil than in the United States. And America can produce passenger aircraft at a lower resource cost than can Brazil.

One country is said to have an **ABSOLUTE ADVANTAGE** over another in the production of a particular good if it can produce that good using smaller quantities of resources than can the other country.

We say that in such a situation Brazil has an **absolute advantage** in coffee production, and the United States has an absolute advantage in aircraft production. And, in such cases, it is obvious that both countries can gain by producing the item in which they have an absolute advantage and then trading with one another.

What is much less obvious—and is in fact one of the great ideas of economics—is that two countries can generally gain from trade *even if one of them is more efficient than the other in producing everything*. A simple parable will help explain why.

Some lawyers are better typists than their secretaries. Should such a lawyer fire her secretary and do her own typing? Not likely. Even though the lawyer may type better than the secretary, good judgment tells her to concentrate on the practice of law and leave the typing to a lower-paid secretary. Why? Because the *opportunity cost* of an hour devoted to typing is an hour less spent practicing law, which is a far more lucrative activity.

One country is said to have a **COMPARATIVE ADVANTAGE** over another in the production of a particular good relative to other goods if it produces that good less inefficiently as compared with the other country.

This is an example of the principle of **comparative advantage** at work. The lawyer specializes in arguing cases despite her absolute advantage in typing because she has a still greater absolute advantage as an attorney. She suffers some direct loss by not doing her own typing. But that loss is more than compensated for by the income she earns selling her legal services to clients.

Precisely the same principle applies to nations, and it underlies the economic analysis of patterns of international trade. The principle, called the *law of comparative advantage*, was discovered by David Ricardo, one of the giants in the history of economic analysis (see the box on the next page). It is one of our **Ideas for Beyond the Final Exam.**

THE LAW OF COMPARATIVE ADVANTAGE

Even if one country is at an absolute disadvantage relative to another country in the production of *every* good, it is said to have a *comparative advantage* in making the good at which it is *least inefficient* (compared with the other country).

Ricardo discovered that two countries can still gain by trading even if one country is more efficient than another in the production of every commodity—that is, has an absolute advantage in every commodity.

In determining the most efficient patterns of production, it is comparative advantage, not absolute advantage, that matters. Thus, a country can gain by importing a good even if that good can be produced at home more efficiently than it can be produced abroad. Such imports make sense if they enable the country to specialize in producing goods at which it is even more efficient.

DAVID RICARDO (1772 1823)

David Ricardo was born 4 years before publication of Adam Smith's *Wealth of Nations*. Descended from a wealthy Jewish family of Portuguese origins, he had about 20 brothers and sisters. Ricardo's formal education ended at the age of 13, and so he was largely self-educated. He began his career by working in his father's stock brokerage firm. At age 21, Ricardo married a Quaker woman and decided to become a Unitarian, a sect then considered "little better than atheist." By Jewish custom, Ricardo's father broke with him, though apparently they remained friendly.

Ricardo then decided to go into the brokerage business on his own and was enormously successful. During the Napoleonic Wars he regularly scored business coups over leading British and foreign financiers, including the Rothschilds. After gaining a huge profit on government securities that he had bought just before the Battle of Waterloo, Ricardo decided to retire from business when he was just over 40 years old.

He purchased a country estate, Gatcomb (now owned by the royal family), where a brilliant group of intellectuals met regularly. Particularly remarkable for the period was the number of women included in the circle, among them Maria Edgeworth, the novelist (who wrote extravagant praise of Ricardo's mind), and Jane Marcet, an author of textbooks, one of which was probably the first textbook in economics. Ricardo's close friends included the economists T. R. Malthus and James Mill, father of John Stuart Mill, the noted philosopher-economist. Malthus remained a close friend of Ricardo even though they disagreed on many subjects and continued their arguments in personal correspondence and in their published works.

James Mill persuaded Ricardo to go into Parliament. As was then customary, Ricardo purchased his seat by buying a piece of land that entitled its owner to a seat in Parliament. There he proved to be a noteworthy advocate of many causes that were against his personal interests.

James Mill also helped persuade Ricardo to write his masterpiece, *The Principles of Political Economy and Taxation*, which may have been the first book of pure economic theory. It was noteworthy that Ricardo, the most practical of practical men, had little patience with empirical economics and preferred instead to rest his analysis explicitly and exclusively on theory. His book made considerable contributions to the analysis of pricing, wage determination, and the effects of various types of taxes, among many other subjects. It also gave us the law of comparative advantage.

Ricardo died in 1823 at the age of 51. He seems to have been a wholly admirable person—honest, charming, witty, conscientious, brilliant—altogether too good to be true.



THE ARITHMETIC OF COMPARATIVE ADVANTAGE

TABLE 34-2

ALTERNATIVE OUTPUTS FROM ONE YEAR OF LABOR INPUT

	In the United States	In Japan
Computers	50	10
Televisions	50	40

Let's see precisely how this works using a hypothetical example that gives a somewhat exaggerated impression of the trading positions of the United States and Japan a few years ago. We imagine that labor is the only input used to produce microcomputers and television sets in the two countries. Suppose further that America has an absolute advantage in both goods, as indicated in Table 34-2. In this example, a year's worth of labor can produce either 50 computers or 50 TV sets in the United States, but only 10 computers or 40 televisions in Japan. So America is the more efficient producer of both goods. Nonetheless, as our lawyer-secretary example suggests, it pays for the United States to specialize and trade with Japan.

To demonstrate this conclusion, we begin by noting that America has a comparative advantage in computers while Japan has a comparative advantage in TVs. Specifically, the numbers in Table 34-2 show that the United States can produce 50 televisions with a year's labor while Japan can produce only 40; so the

TABLE 34-3 EXAMPLE OF THE GAINS FROM TRADE

	United States	Japan	Total
Computers (thousands)	+25	−10	+15
Televisions (thousands)	−25	+40	+15

United States is 25 percent more efficient than Japan in producing TV sets. However, America is five times as efficient as Japan in producing computers: It can produce 50 per year of labor rather than 10. Because America's competitive edge is far greater in computers than in televisions, we say that the United States has a *comparative advantage* in computers.

From the Japanese perspective, these same numbers indicate that Japan is only slightly less efficient than America in TV production but drastically less efficient in computer production. So Japan's comparative advantage is in the television industry. According to Ricardo's law of comparative advantage, then, the two countries can gain if the United States specializes in producing computers, Japan specializes in producing TVs, and the two countries trade. Let us verify that this is true.

Suppose Japan transfers 1,000 years of labor out of the computer industry and into TV manufacturing. According to the figures in Table 34-2, its computer output falls by 10,000 units while its TV output rises by 40,000 units. (See the middle column of Table 34-3.) Suppose, at the same time, the United States transfers 500 years of labor out of television manufacturing (thereby losing 25,000 TVs) and into computer making (thereby gaining 25,000 computers). Table 34-3 shows us that these transfers of resources in the two countries increase the world's production of both outputs. Together, the two countries now have 15,000 additional TVs and 15,000 additional computers—surely a nice outcome.

Was there some sleight of hand here? How did both the United States and Japan gain both computers and TVs? The explanation is that the process we have just described involves more than just a swap of a fixed bundle of commodities. It is also a *change in the production arrangements*, with some of Japan's inefficient computer production taken over by more efficient American makers, and with some of America's TV production taken over by Japanese television companies who are *less* inefficient at making TVs than Japanese computer manufacturers are at making computers. The underlying principle is both simple and fundamental:

When every country does what it can do best, all countries can benefit because more of every commodity can be produced without increasing the amounts of labor used.

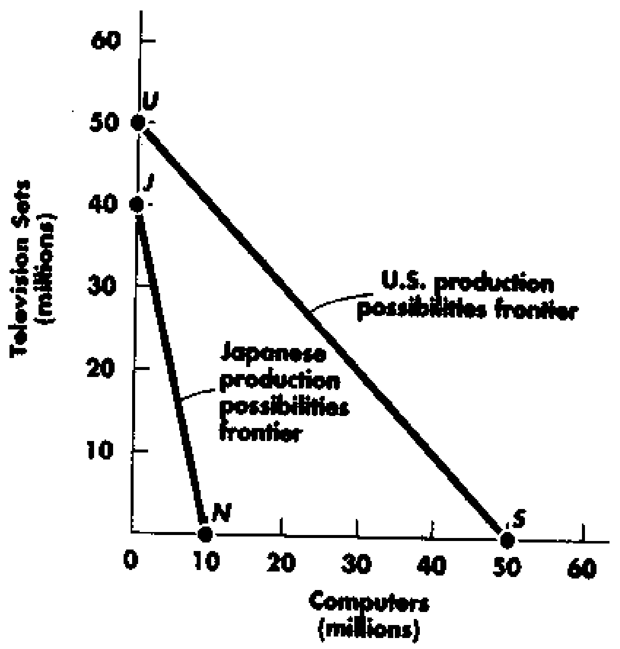
THE GRAPHICS OF COMPARATIVE ADVANTAGE

The gains from trade can also be displayed graphically, and doing so helps us understand whether they are large or small.

The lines *US* and *JN* in Figure 34-1 are closely related to the production possibilities frontiers of the two countries, but differ in that they pretend that each

FIGURE 34-1 PER-CAPITA PRODUCTION POSSIBILITIES FRONTIERS FOR TWO COUNTRIES

Because America can produce more of every commodity using the same quantity of labor as does Japan, its per-capita production possibilities frontier, *US*, is higher than Japan's, *JN*. But the United States has a comparative advantage in computers, in which it is 5 times as productive as Japan. (It can produce 50 million computers, point *S*, compared with Japan's 10 million, point *N*.) On the other hand, America is only 25 percent more productive in TVs (point *U*) than Japan (point *J*). Thus, Japan has a comparative advantage in producing televisions.



country has the same amount of labor available—in this case, 1 million person-years.[1] For example, Table 34-2 tells us that, for each 1 million person-years of labor, the United States can produce 50 million TVs and no computers (point *U* in Figure 34-1), 50 million computers and no TVs (point *S*), or any combination between (the line *US*). Similar reasoning leads to line *JN* for Japan.

America's actual production possibilities frontier would be even higher, relative to Japan's, than shown in Figure 34-1 because the U.S. population is larger. But Figure 34-1 is more useful to us because it highlights the differences in efficiency that determine both absolute and comparative advantage. Let us see how.

The fact that line *US* lies *above* line *JN* means that the United States can manufacture more televisions and more computers than Japan with the same amount of labor. This reflects our assumption that America has an *absolute* advantage in both commodities.

America's comparative advantage in computer production and Japan's comparative advantage in TV production are shown in a different way—by the relative *slopes* of the two lines. Look back to Table 34-2, which shows that the United States can acquire a computer on its own by giving up a TV. Thus, the opportunity cost of a computer in the United States is one television set. This opportunity cost is depicted graphically by the slope of the U.S. production possibilities frontier in Figure 34-1, which is $OU/OS = 50/50 = 1$.

Table 34-2 also tells us that the opportunity cost of a computer in Japan is four TVs. And this is depicted in Figure 34-1 by the slope of Japan's production possibilities frontier, which is $OJ/ON = 40/10 = 4$.

A country's absolute advantage in production over another country is shown by its having a higher per-capita production possibilities frontier. The difference in

[1] To review the concept of the production possibilities frontier, see Chapter 3.

the comparative advantages of the two countries is shown by the difference in the slopes of their frontiers.

Because opportunity costs differ in the two countries, gains are possible if the two countries specialize and trade with one another. Specifically, it is cheaper, in terms of real resources forgone, for *either* country to acquire its computers in the United States. By a similar line of reasoning, the opportunity cost of TVs is higher in the United States than in Japan, so it makes sense for both countries to acquire their televisions in Japan.[2]

Notice that if the slopes of the two production possibilities frontiers, *JN* and *US*, were equal, then opportunity costs would be the same in each country. In that case, there would be no potential gains from trade. Gains from trade arise from *differences* across countries, not from similarities. This is an important point on which people are often confused. It is often argued that two very different countries—say, the United States and Mexico—cannot gain much by trading with one another. In fact:

Two very similar countries may gain little from trade. Large gains from trade are most likely when countries are very different.

How the gains from trade are divided between the two countries depends on the prices that emerge from world trade, which is the subject of the next section. But we already know enough to see that world trade must, in our example, leave a computer costing more than one TV and less than four. Why? Because, if a computer brought less than one TV (its opportunity cost in the United States) on the world market, America would produce its own TVs rather than buying them from Japan. And if a computer cost more than four TVs (its opportunity cost in Japan), Japan would prefer to produce its own computers rather than buy them from the United States.

We conclude, therefore, that if both countries are to trade, the rate of exchange between TVs and computers must be somewhere between 4 to 1 and 1 to 1. To illustrate the gains from trade in a concrete example, suppose the world price ratio settles at 2 to 1; that is, one computer costs the same as two televisions. How much, precisely, do the United States and Japan gain from world trade?

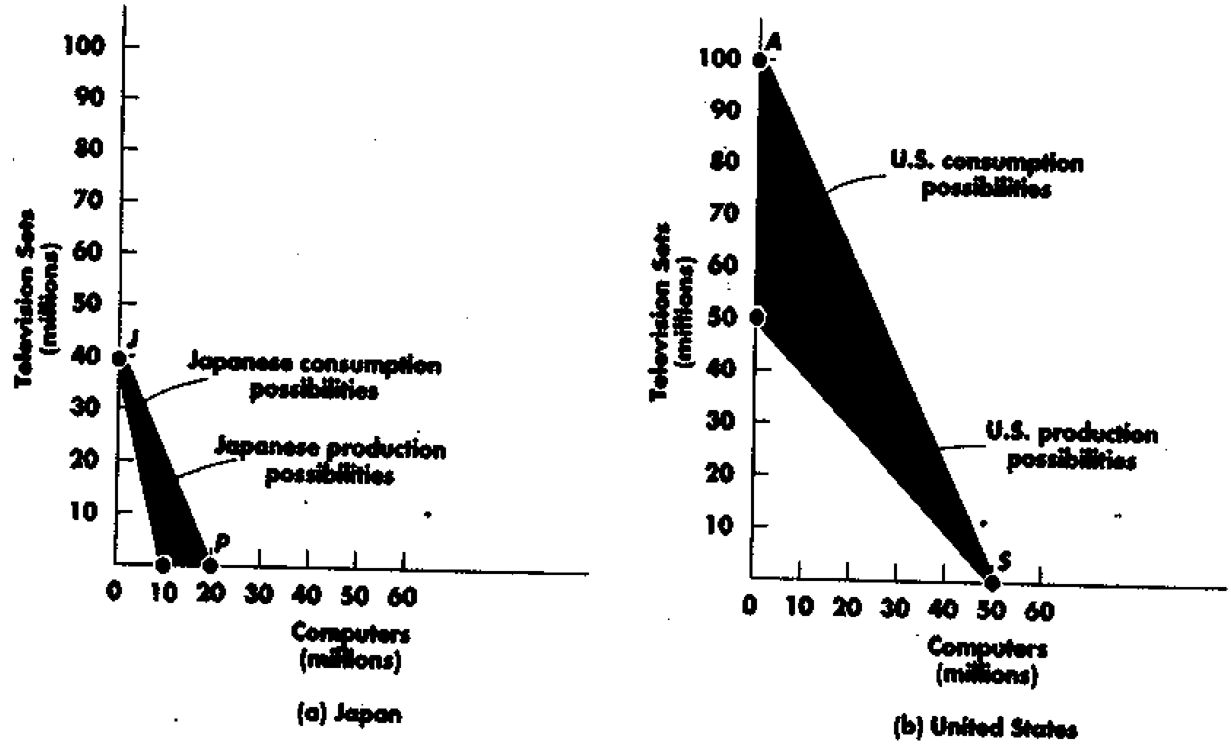
Figure 34-2 is designed to help us see the answer. Production possibilities frontiers *US* in Panel (b) and *JN* in Panel (a) are the same as in Figure 34-1. But America can do better than line *US*. Specifically, with a world price ratio of 2 to 1, the United States can buy a TV by giving up only one-half of a computer, rather than one (which is the opportunity cost of TVs in America). Hence, if the United States produces only computers [point *S* in Figure 34-2(b)] and buys its TVs from Japan, America's *consumption possibilities* will be as indicated by the blue line that begins at point *S* and has a slope of 2—indicating that each computer it sells brings America two television sets. Since trade allows the United States to choose a point on *AS* rather than on *US*, trade opens up consumption possibilities that were simply not available before.

The story is similar for Japan. If the Japanese produce only televisions [point *J* in Figure 34-2(a)], they can acquire a computer from the United States for every two TVs they give up as they move along the blue line *JP* (whose slope is 2). This is better than they can do on their own, since a sacrifice of two TVs in Japan yields only one-half of a computer. Hence, world trade enlarges Japan's consumption possibilities from *JN* to *JP*.

---

[2]As an exercise, provide this line of reasoning.

FIGURE 34-2 THE GAINS FROM TRADE



In this diagram, we suppose that trade opens up between the United States and Japan and that the world price of computers is twice the world price of TVs. Now America's consumption possibilities are all the points on line *AS* (which starts at *S* and has a slope of 2), rather than just the points on its own production possibilities frontier, *US*. Similarly, Japan can choose any point on line *JP* (which begins at *J* and has a slope of 2), rather than just points on *JN*. Thus, both nations gain from trade.

Figure 34-2 shows graphically that gains from trade arise to the extent that world prices (2 to 1 in our example) differ from domestic opportunity costs (4 to 1 and 1 to 1 in our example). So it is a matter of some importance to understand how prices in international trade are established. This we shall do shortly.

COMPARATIVE ADVANTAGE AND COMPETITION OF "CHEAP FOREIGN LABOR"

But first let us observe that the principle of comparative advantage takes us a long way toward uncovering the fallacy in the "cheap foreign labor" argument described earlier in the chapter. Given the assumed productive efficiency of American labor, and the inefficiency of Japanese labor, we would expect wages to be much higher in the United States. And, indeed, they were until recent years.

In these circumstances, one might expect American workers to be apprehensive about an agreement to permit open trade between the two countries—"How can we hope to meet the unfair competition of those underpaid Japanese workers?" And Japanese laborers might also be concerned—"How can we hope to meet the competition of those Americans, who are so efficient in producing everything?"

The principle of comparative advantage shows us that both fears are unjustified. As we have just seen, when trade is opened up between Japan and the

United States, *workers in both countries will be able to earn higher real wages than before* because of the increased productivity that comes about through specialization.

As Figure 34-2 shows, once trade opens up, Japan can end up with more TVs and more computers than it had before. So the living standards of its workers can rise even though they have been left vulnerable to the competition of the superefficient Americans. The United States also can end up with more TVs and with more computers; so the living standards of its workers can rise even though they have been exposed to the competition of cheap Japanese labor. These higher standards of living are, of course, a reflection of the higher real wages earned by workers in both countries.

The lesson to be learned here is elementary: Nothing helps raise standards of living more than does a greater abundance of goods.

SUPPLY-DEMAND EQUILIBRIUM AND PRICING IN WORLD TRADE

How the gains from trade are shared depends on the prices that emerge from world trade. As usual, price determination in a free market depends on supply and demand.

When applied to international trade, however, the supply-demand model runs into several new complications. First, it involves at least two demand curves: that of the exporting country and that of the importing country. Second, it may also involve two supply curves, since the importing country may produce part of its own consumption. Third, equilibrium does not take place at the intersection point of *either* pair of supply-demand curves. Why? Because if there is any trade, the exporting country's quantity supplied must be *greater* than its quantity demanded, while the quantity supplied by the importing country must be *less* than its quantity demanded.

These complications are illustrated in Figure 34-3, where we show the supply and demand curves of a country that exports wheat in Panel (a) and of a country that imports wheat in Panel (b). For simplicity, we assume that these countries do not deal in wheat with anyone else.

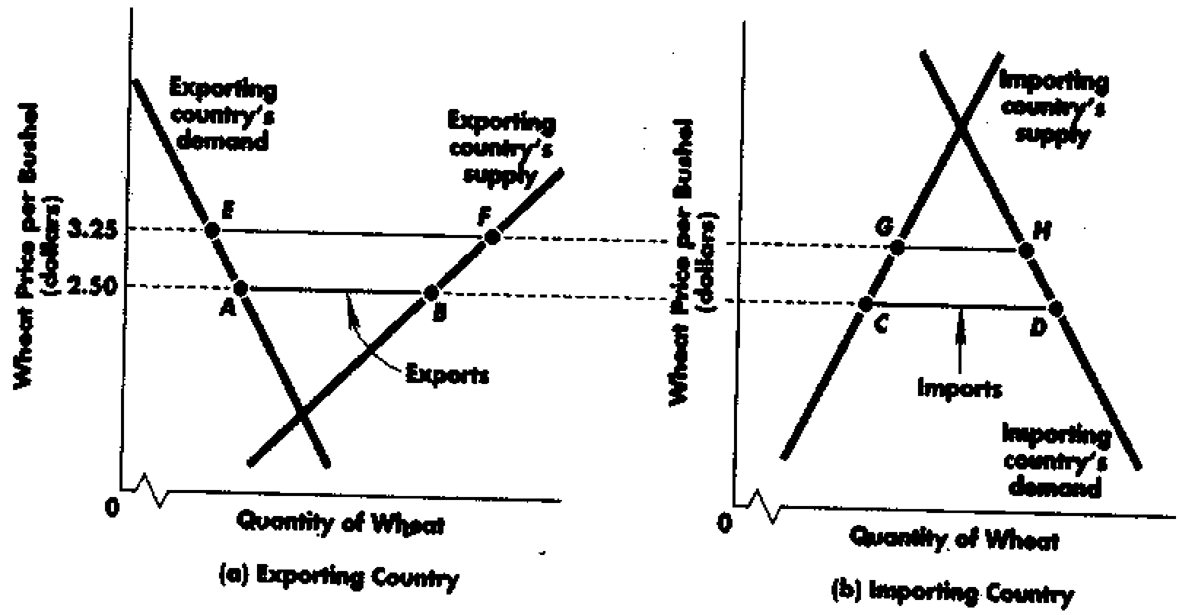
Where will the two-country wheat market reach equilibrium? If there is free trade, the equilibrium price must satisfy two requirements:

1. The price of wheat must be the same in both countries.
2. The quantity of wheat exported must equal the quantity of wheat imported.

In Figure 34-3, this happens at a price of $2.50 per bushel. At that price, the distance *AB* between what the exporting country produces and what it consumes equals the distance *CD* between what the importing country consumes and what it produces. This means that, at a price of $2.50 per bushel, the amount the exporting country wants to sell is exactly equal to the amount the importing country wants to buy.

At any price above $2.50, producers in both countries will want to sell more and consumers in both countries will want to buy less. For example, if the price rises to $3.25 per bushel, the exporter's quantity supplied will rise from *B* to *F*, and the exporter's quantity demanded will fall from *A* to *E*, as shown in Figure 34-3(a). As a result, there will be more available for export—*EF* rather than *AB*. For exactly the same reason, the price increase will cause higher production and lower sales in the importing country, leading to a reduction in imports from *CD* to *GH* in Panel (b).

FIGURE 34-3 SUPPLY-DEMAND EQUILIBRIUM IN THE INTERNATIONAL WHEAT TRADE



Equilibrium requires that exports, *AB* (which is the exporting nation's quantity supplied, *B*, minus the exporter's quantity demanded, *A*), exactly balance imports, *CD*, by the importing country. At \$2.50 per bushel of wheat, there is equilibrium. But at a higher price, say, \$3.25, there is disequilibrium because export supply, *EF*, exceeds import demand, *GH*.

But this means that the higher price, \$3.25 per bushel, cannot be sustained in a free and competitive international market. With export supply *EF* far greater than import demand *GH*, there must be downward pressure on price and a move back toward the \$2.50 equilibrium price. Similar reasoning shows that prices below \$2.50 also cannot be sustained. Thus:

In international trade, the equilibrium price is the one that makes the exporting country want to export exactly the amount that the importing country wants to import. Equilibrium will thus occur at a price at which the horizontal distance *AB* in Figure 34-3(a) (the excess of the exporter's quantity supplied over its quantity demanded) is equal to the horizontal distance *CD* in Figure 34-3(b) (the excess of the importer's quantity demanded over its quantity supplied). At this price, the *world's* quantity demanded is equal to the *world's* quantity supplied.

TARIFFS, QUOTAS, AND OTHER INTERFERENCES WITH TRADE

Despite the mutual gains from international trade, nations often interfere with the operation of free international markets. In fact, until the rise of the free-trade movement about 200 years ago (with such economists as Adam Smith and David Ricardo as its vanguard), it was taken for granted that one of the essential tasks of government was to impede trade—presumably in the national interest.

There were many who argued then (and many who still argue today) that the proper aim of government policy was to promote exports and discourage imports, for that would increase the amount foreigners owed the nation. According to this view, a nation's wealth consists of the amount of gold or other monies at its command.

Obviously, there are limits to which such a policy can be pursued. A country *must* import vital foodstuffs or critical raw materials that it cannot supply for itself. Moreover, it is mathematically impossible for *every* country to sell more than it buys, for one country's exports *must* be some other country's imports. If everyone competes in this game and cuts imports to the bone, then obviously exports must go the same way. The result will be that everyone is deprived of the mutual gains from trade—which is precisely what happens in a trade war.

After the protectionist 1930s, the United States moved away from policies designed to impede imports and gradually assumed a leading role in promoting free trade. Over the last 50 years, tariffs and other trade barriers have been reduced dramatically. In recent years, the United States has led the world to complete the Uruguay Round of tariff reductions, and has joined Canada and Mexico in the North American Free Trade Agreement (NAFTA). The latter caused a political firestorm in the United States in 1993 and 1994, with critic Ross Perot predicting a "giant sucking sound" as American workers lost their jobs to competition from "cheap Mexican labor." (Does that sound familiar?)

Modern governments use three main devices when seeking to control trade: tariffs, quotas, and export subsidies.

A **TARIFF** is a tax on imports.

A **tariff** is simply a tax on imports. An importer of wheat, for example, may be charged $1 for each bushel brought into the country. The United States is generally a low-tariff country, with only a few notable exceptions such as the 25 percent tariff on light trucks. However, many other countries rely on heavy tariffs to protect their industries. Tariff rates of 100 percent or more are not uncommon.

A **QUOTA** specifies the maximum amount of a good that is permitted into the country from abroad per unit of time.

A **quota** is a legal limit on the amount of a good that may be imported. For example, the government might allow no more than 25 million bushels of wheat to be imported in a year. In some cases, governments ban the importation of certain goods outright—a quota of zero. The United States now imposes quotas on a smattering of goods, including textiles, meat, and sugar. (See the box on the next page.) But most imports are free of quotas.

An **EXPORT SUBSIDY** is a payment by the government to exporters to permit them to reduce the selling prices of their goods so they can compete more effectively in foreign markets.

An **export subsidy** is a payment by the government to an exporter. By reducing the exporter's costs, such subsidies permit exporters to lower their selling prices and to compete more effectively in world trade. While export subsidies are minor in the United States, they are used extensively by some foreign governments to assist their industries—a practice that provokes bitter complaints from American manufacturers about "unfair competition." For example, years of heavy government subsidies helped the European Airbus consortium take a sizable share of the world market for commercial aircraft away from American manufacturers.

HOW TARIFFS AND QUOTAS WORK

Both tariffs and quotas restrict supplies coming from abroad and drive up prices. A tariff works by raising prices and hence cutting the demand for imports, while the sequence associated with a quota is just the reverse—a restriction in supply forces prices up.

The supply and demand curves in Figure 34-4 illustrate how tariffs and quotas work. Just as in Figure 34-3, the equilibrium price of wheat under free trade is $2.50 per bushel (in both countries). At this price, the exporting country produces 125 million bushels [point *B* in Panel (a)] and consumes 80 million (point *A*); so its exports are 45 million bushels—the distance *AB*. Similarly, the importing country consumes 95 million bushels [point *D* in Panel (b)] and produces

HOW SWEET IT IS: THE U.S. SUGAR QUOTA



The United States has restricted sugar imports since 1934. But the current program, which many economists feel is the most egregious of all our agricultural subsidies, dates only from 1990. It is called a "tariff-rate quota," which most economists view as a fancy name for grotesque protectionism.

Each year a small amount of imported sugar is allowed to enter the United States under a negligible tariff. Beyond that, however, the U.S. government posts a tariff so high that no one wants to pay it in order to import sugar. (Lately, the high tariff has been 15 cents per pound, which is generally above the world price of raw sugar.) In consequence, our domestic price of sugar is roughly twice the world price.

This high price is, of course, a great boon to American producers of sugar and sugar-substitutes like corn syrup. But it costs U.S. consumers of sugar, and manufacturers of candy and soft drinks, between $1 billion and $2 billion per year. And it has also led to some bizarre attempts to evade the quota—like importing sugar as an ingredient in chocolate and Kool-Aid, and then removing the sugar content in the United States.

only 50 million (Panel C), so that its imports are also 45 million bushels—the distance *CD*.

Now suppose the government of the importing nation imposes an import quota of (no more than) 30 million bushels. The free-trade equilibrium is no longer possible. Instead, the market must equilibrate at a point where both exports and imports are 30 million bushels. As Figure 34-4 indicates, this requires different prices in the two countries.

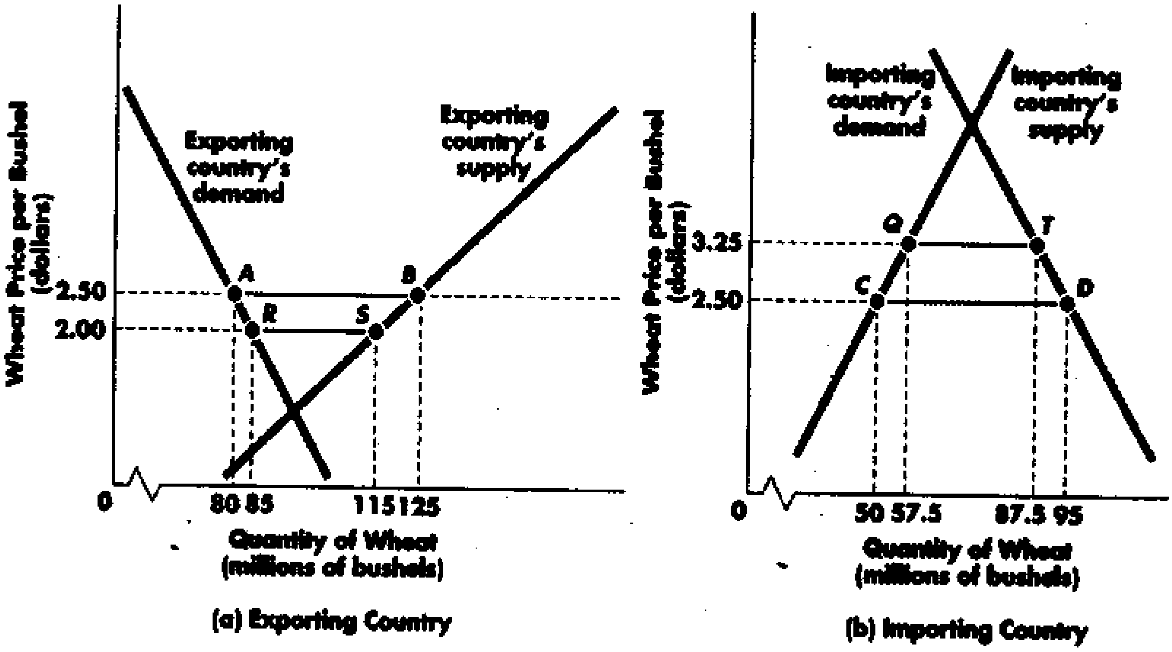
Imports in Panel (b) will be 30 million—the distance *QT*—only when the price of wheat in the importing nation is $3.25 per bushel, because only at this price will quantity demanded exceed domestic quantity supplied by 30 million bushels. Similarly, exports in Panel (a) will be 30 million bushels—the distance *RS*—only when the price in the exporting country is $2.00 per bushel. At this price, quantity supplied exceeds domestic quantity demanded by 30 million bushels in the exporting country. Thus, the quota *raises* the price in the importing country to $3.25 and *lowers* the price in the exporting country to $2.00. In general:

An import quota on a product normally will reduce the volume of that product traded, raise the price in the importing country, and reduce the price in the exporting country.

The same restriction of trade can be accomplished through a tariff. In our example, a quota of 30 million bushels led to a price that was $1.25 higher in the importing country than in the exporting country ($3.25 versus $2.00). Suppose that, instead of a quota, the importing nation posts a $1.25 per bushel tariff. International trade equilibrium then must satisfy the following two requirements:

1. The price that consumers in the importing country pay for wheat must exceed the price that suppliers in the exporting country receive by $1.25 (the amount of the tariff).
2. The quantity of wheat exported must equal the quantity of wheat imported.

FIGURE 34-4 QUOTAS AND TARIFFS IN INTERNATIONAL TRADE



Under free trade, the equilibrium price of wheat is $2.50 per bushel. The exporting country, in Panel (a), sends *AB*, or 45 million bushels, to the importing country (distance *CD*). If a quota of 30 million bushels is imposed by the importing country, these two distances must shrink to 30 million bushels. The solution is shown by distance *RS* for exports and distance *QT* for imports. Exports and imports are equal, as must be the case, but the quota forces prices to be unequal in the two countries. Wheat sells for $3.25 per bushel in the importing country but only $2.00 per bushel in the exporting country. A tariff achieves the same result differently. It *requires* that the prices in the two countries be $1.25 apart. And this, as the graph shows, dictates that exports and imports will be equal at 30 million bushels.

By consulting the graphs in Figure 34-4, you can see exactly where these two requirements are satisfied. If the exporter produces at *S* and consumes at *R*, while the importer produces at *Q* and consumes at *T*, then exports and imports are equal (at 30 million bushels), and the two domestic prices differ by exactly $1.25. (They are $3.25 and $2.00.) What we have just discovered is a general result of international trade theory:

Any restriction of imports that is accomplished by a quota normally can also be accomplished by a tariff.

In this case, the tariff corresponding to an import quota of 30 million bushels is $1.25 per bushel.

■ TARIFFS VERSUS QUOTAS

But while tariffs and quotas can accomplish the same reduction in international trade and lead to the same domestic prices in the two countries, there *are* some important differences between the two types of restrictions.

First, under a quota, profits from the price increases in the importing country usually go into the pockets of the foreign and domestic sellers of the product.

Because supplies are limited by quotas, customers in the importing country must pay more for the product. So the suppliers, whether foreign or domestic, receive more for every unit they sell. For example, it has been estimated that U.S. import quotas on Japanese cars in the early 1980s raised the profits of both American and Japanese automakers by billions of dollars per year.

On the other hand, when trade is restricted by a tariff, some of the profits go instead as tax revenues to the *government* of the importing country. In effect, the government increases its tax revenues partly at the expense of its citizens and partly at the expense of foreign exporters, who must accept a reduced price because of the resulting decrease in quantity demanded in the importing country. (Domestic producers again benefit, because they are exempt from the tariff.) In this respect, a tariff is certainly a better proposition than a quota from the viewpoint of the country that enacts it.

Another important distinction between the two measures is their different implications for productive efficiency. A tariff handicaps all foreign suppliers equally. It still awards sales to the firms and nations that are most efficient and can therefore supply the goods most cheaply. A quota, on the other hand, necessarily awards its import licenses more or less capriciously—perhaps on a first-come, first-served basis or in proportion to past sales or even on political criteria. There is not the slightest reason to expect the most efficient suppliers to get the import permits.

The U.S. quota on Japanese cars in the 1980s illustrates all of these effects. Japanese automakers responded to the limit on the number of cars by shipping bigger models equipped with more "optional" equipment. The "stripped-down" Japanese car became a thing of the past. And the newer, smaller Japanese automakers—like Subaru and Mitsubishi—found it difficult at first to compete in the U.S. market because their quotas were so much smaller than those of Toyota, Nissan, and Honda.

If a country must inhibit imports, there are two important reasons for it to give preference to tariffs over quotas: (1) Some of the resulting financial gains from tariffs go to the government of the importing country rather than to foreign and domestic producers; and (2) unlike quotas, tariffs offer special benefits to more efficient exporters.

WHY INHIBIT TRADE?

To state that tariffs are a better way to inhibit international trade than quotas leaves open a far more basic question: Why limit trade in the first place? It has been estimated that trade restrictions cost American consumers about $70 billion per year in the form of higher prices. Why should they be asked to pay these higher prices?

Two answers are commonly given. One holds that a country should restrict trade to get more advantageous prices for its own goods. The other argues that particular industries need protection from foreign competition. Let us examine each in turn.

GAINING A PRICE ADVANTAGE

A tariff forces foreign exporters to sell more cheaply by restricting their market. If they do not cut their prices, they will be left with unsold goods. Suppose, as in Figure 34-4(b), that a $1.25 tariff on wheat raises the price of a bushel in the

importing country from $2.50 to $3.25 per bushel. This rise in price drives down imports from an amount represented by the length of the orange line *CD* to the smaller amount represented by the blue line *QT*. And to the exporting country, this means an equal reduction in exports [see the change from *AB* to *RS* in Figure 34-4(a)].

As a result, the price at which the exporting country can sell its wheat is driven down (from $2.50 to $2.00 in the example) while producers in the importing country—being exempt from the tariff—can charge $3.25 per bushel. In effect, such a tariff amounts to government intervention to rig prices in favor of domestic producers.

However, this technique works only as long as foreigners accept tariff exploitation passively. They rarely do. Instead, they retaliate, usually by imposing tariffs or quotas of their own on their imports from the country that first began the tariff game. This can easily lead to a trade war in which no one gains in terms of more favorable prices and everyone loses in terms of the resulting reductions in overall trade. Something like this happened to the world economy in the 1930s and helped prolong the worldwide depression.

Tariffs or quotas can benefit particular domestic industries in a country that is able to impose them without fear of retaliation. But when every country uses them, everyone is likely to lose in the long run.

PROTECTING PARTICULAR INDUSTRIES

The second, and probably more frequent, reason why countries restrict trade is to protect particular industries from foreign competition. If foreigners can produce steel or watches or shoes more cheaply, domestic businesses and unions in these industries are quick to demand protection; and their governments are often reluctant to deny it to them. It is here that the cheap foreign labor argument is most likely to be invoked.

Protective tariffs and quotas are explicitly designed to rescue firms whose relative inefficiency does not permit them to compete with foreign exporters in an open, world market. But it is precisely the harsh competition from abroad that gives consumers the benefits of international specialization. In our numerical example of comparative advantage, one can well imagine the complaints from Japanese computer makers as the opening of trade led to increased importation of U.S. computers. At the same time, American TV manufacturers would probably have expressed outrage over the flood of imported TVs from Japan. Yet it is Japanese specialization in televisions and U.S. specialization in computers that enables citizens of both countries to enjoy higher standards of living. If governments interfere with this process, consumers in both countries will lose out.

Industries threatened by foreign competition often argue that some form of protection against imports is needed to prevent loss of jobs. This argument is frequently made, for example, by the U.S. apparel industry since nearly two-thirds of all clothing purchased by Americans is imported. But we know from basic macroeconomics that there are better ways to stimulate employment, such as raising aggregate demand.

A program that limits foreign competition will be more effective at preserving employment *in the particular protected industry*. But it will typically do so at a high cost to consumers and to the economy. Table 34-4 gives some estimates of the costs to American consumers of using tariffs and quotas to save jobs in selected industries. In every case, the costs far exceed the wages of the workers in the protected industries—ranging as high as $600,000 per job for the sugar quota.

TABLE 34-4 ESTIMATED COSTS OF PROTECTIONISM TO CONSUMERS

Industry	Cost per Job Saved
Apparel	$139,000
Costume jewelry	97,000
Shipping	415,000
Sugar	600,000
Textiles	202,000
Women's footwear	102,000

SOURCE: Gary C. Hufbauer and Kimberly Ann Elliott, *Measuring the Costs of Protectionism in the United States* (Washington, D.C.: Institute for International Economics, January 1994), Table 1.3, pp. 12–13.

Nevertheless, complaints over proposals to reduce a tariff or a quota are justified unless something is done to ease the cost to individual workers of switching to the lines of production that trade makes profitable.

The argument for free trade between countries cannot be considered airtight if there is no adequate program to assist those few citizens in each country who are harmed whenever patterns of production change drastically—as would happen, for example, if tariff and quota barriers were suddenly reduced.

Owners of television factories in the United States and of computer factories in Japan may see heavy investments suddenly rendered unprofitable. So would workers whose investments in acquiring special skills and training are no longer marketable. Nor are the costs to displaced workers only monetary. They may have to move to new locations as well as to new industries, uprooting their families, losing old friends and neighbors, and so on. That the *majority* of citizens undoubtedly gain from free trade is no consolation to those who are its victims.

To mitigate this problem, the United States follows two basic approaches. First, provisions in the trade laws offer temporary protection from sudden surges of imports, on the grounds that unexpected changes in trade patterns do not give people enough time to adjust.

TRADE ADJUSTMENT ASSISTANCE provides special unemployment benefits, loans, retraining programs, and other aid to workers and firms that are harmed by foreign competition.

Second, the government has set up **trade adjustment assistance** programs to help workers and businesses who lose their jobs or their markets because of increases in imports. Firms may be eligible for technical assistance, government loans or loan guarantees, and permission to delay tax payments. Workers may qualify for retraining programs, longer periods of unemployment compensation, and funds to defray moving costs. Each form of assistance is designed to ease the burden on the victims of free trade so that the rest of us can enjoy its considerable benefits.

OTHER ARGUMENTS FOR PROTECTION

NATIONAL DEFENSE AND OTHER NONECONOMIC CONSIDERATIONS

Sometimes a tariff or other interference with trade may be justified on noneconomic grounds. If a country considers itself vulnerable to military attack, it may

be perfectly rational to keep alive industries whose outputs can be obtained more cheaply abroad but whose supplies might be cut off in an emergency. The argument is valid, but there is a danger: Industries with the most peripheral relationship to defense are likely to invoke this argument on their behalf. For instance, for years the U.S. watchmaking industry claimed protection for itself on the grounds that its skilled craftsmen would be invaluable in wartime.

Similarly, the United States has occasionally banned either exports to or imports from nations such as Cuba, Iran, and Iraq on political grounds. Such actions often have important economic effects, creating either bonanzas or disasters for particular American industries. But they are justified by politics, not by economics. Noneconomic reasons also explain quotas on importation of whaling products and on the furs of other endangered species.

THE INFANT-INDUSTRY ARGUMENT

Another common argument for protectionism is the so-called *infant-industry argument*. Promising new industries, it is alleged, often need breathing room to flourish and grow. If we expose these infants to the rigors of international competition too soon, the argument goes, they may never develop to the point where they can survive on their own in the international marketplace.

The argument, while valid in certain instances, is less defensible than it seems at first. It makes sense to protect an infant industry only if the prospective future gains are sufficient to repay the social losses incurred while it is being protected. But if the industry is likely to be so profitable in the future, why doesn't private capital rush in to take advantage of the prospective net profits? The annals of business are full of cases in which a new product or a new firm lost money at first but profited handsomely later.

The infant-industry argument for protection stands up to scrutiny only where funds are not available to a particular industry for some reason, despite its glowing profit prospects. And even then it may make more sense to provide a government loan to the infant industry than to provide trade protection.

It is hard to think of examples where the infant-industry argument applies. But even if such a case were found, one would have to be careful that the industry not remain in diapers forever. There are too many cases in which new industries were awarded protection when they were being established and, somehow, the time to withdraw that protection never arrived. One must beware of infant industries that never grow up.

STRATEGIC TRADE POLICY

A stronger argument for (temporary) protection is beginning to have substantial influence on U.S. trade policy. Advocates of this argument, including some top officials in the Clinton administration, agree that free trade for all is the best system. But they point out that we live in an imperfect world in which many nations refuse to play by the rules of the free-trade game. And they fear that a nation that pursues free trade in a protectionist world is likely to lose out. It therefore makes sense, they argue, to *threaten* to protect your markets unless other nations agree to open theirs. (See the article on the next page by columnist William Safire.)

The United States has followed this strategy in trade negotiations with several countries in recent years. The most prominent case came in 1995, when the U.S. government threatened to impose 100 percent tariffs on imported Japanese luxury cars unless Japan agreed to open up its domestic markets for automobiles

CAN PROTECTIONISM SAVE FREE TRADE?

In this 1983 column, William Safire shook off his longstanding attachment to free trade and argued eloquently for retaliation against protectionist nations.

WASHINGTON—Free trade is economic motherhood. Protectionism is economic evil incarnate.... Never should government interfere in the efficiency of international competition.

Since childhood, these have been the tenets of my faith. If it meant that certain businesses in this country went belly-up, so be it.... If it meant that Americans would be thrown out of work by overseas companies paying coolie wages, that was tough....

The thing to keep in mind, I was taught, was the Big Picture and the Long Run. America, the great exporter, had far more to gain than to lose from free trade; attempts to protect inefficient industries here would ultimately cost more American jobs.

While playing with my David Ricardo doll and learning nursery rhymes about comparative advantage, I was listening to another laissez-fairy tale: Government's role in the world of business should be limited to keeping business honest and competitive. In God we antitrusted. Let businesses operate in the free marketplace.

Now American businesses are no longer competing with foreign companies. They are competing with foreign governments who help their local businesses. That means the world arena no longer offers a free marketplace; instead, most other governments are pushing a policy that can be called *helpfulism*.

Helpfulism works like this: A government like Japan decides to get behind its baseball-bat industry. It pumps in capital, knocks off marginal operators, finds subtle ways to discourage imports of Louisville Sluggers, and selects target areas for export blitzes. Pretty soon, the favored Japanese companies are driving foreign competitors batty.

How do we compete with helpfulism? One way is to complain that it is unfair; that draws a horselaugh. Another way is to demand a "Reagan Round" of trade negotiations under GATT, the Gentlemen's Agreement To Talk, which is equally laughable. Yet another way is to join the helpfuls by subsidizing our exports and permitting our companies to try monopolistic tricks abroad not permitted at home. But all that makes us feel guilty, with good reason.

The other way to deal with helpfulism is through—here comes the dreadful word—*protection*. Or, if you prefer a euphemism, *retaliation*. Or if that is still too severe, *reciprocity*. Whatever its name, it is a way of saying to the cutthroat cartelists we sweetly call our trading partners: "You have bent the rules out of shape. Change your practices to conform to the agreed-upon rules, or we will export a taste of your own medicine."

A little balance, then, from the free trade theorists. The demand for what the Pentagon used to call "protective reaction" is not demagoguery, not shortsighted, not self-defeating. On the contrary, the overseas pirates of protectionism and exemplars of helpfulism need to be taught the basic lesson in trade, which is: tit for tat.

SOURCE: William Safire, "Smoot-Hawley Lives," *The New York Times*, March 17, 1983. Copyright © 1983 by The New York Times Company. Reprinted by permission.



and auto parts. A dangerous trade war was narrowly averted when an agreement was reached at the 11th hour.

The strategic argument for protection is a hard one for economists to deal with. While it accepts the superiority of free trade, it argues that threatening protectionism is the best way to establish free trade. Such a strategy might work, but it clearly involves great risks. If threats that America will turn protectionist induce other countries to scrap existing protectionist policies, then the gamble will have succeeded. But, if the gamble fails, protectionism increases.

CAN CHEAP IMPORTS BE BAD FOR A COUNTRY?

DUMPING means selling goods in a foreign market at lower prices than those charged in the home market.

One of the most curious features of the protectionist position is the fear of low prices charged by foreign sellers. Countries that subsidize exports are accused of **dumping**—of getting rid of their goods at unconscionably low prices. For example, Japan and Korea have frequently been accused of dumping a variety of goods on the U.S. market.

Economists find this argument strange. As a nation of consumers, we should be indignant when foreigners charge us *high* prices, not *low* ones. That is the common-sense rule that guides every consumer's daily life. Only from the topsy-turvy viewpoint of an industry seeking protection from competition are low prices seen as being against the public interest.

Ultimately, it must be in the best interest of a country to get its imports as cheaply as possible. It would be ideal for the United States if the rest of the world were willing to provide its exports to us free or virtually so. We could then live in luxury at the expense of the rest of the world.

But, of course, what benefits the United States as a whole does not necessarily benefit every single American. If quotas on, say, sugar imports were dropped, American consumers and industries that purchase sugar would gain from lower prices. But owners of sugar fields and their employees would suffer serious losses in the form of lower profits, lower wages, and lost jobs—losses they would fight hard to prevent. For this reason, politics often leads to the adoption of protectionist measures that would likely be rejected on strictly economic criteria.

CONCLUSION: A LAST LOOK AT THE "CHEAP FOREIGN LABOR" ARGUMENT

The preceding discussion reveals the fundamental fallacy in the argument that American workers have to fear cheap foreign labor. If workers in other countries are willing to supply their products to us with little compensation, this must ultimately *raise* the standard of living of the average American worker. As long as the government's monetary and fiscal policies succeed in maintaining high levels of employment at home, how can we possibly lose by getting the products of the world at bargain prices?

There are, however, some important qualifications. First, our macroeconomic policy may not be effective. If workers who are displaced by foreign competition cannot find jobs in other industries, then American workers will indeed suffer from international trade. But that is a shortcoming of the government's monetary and fiscal policies, not of its international trade policies.

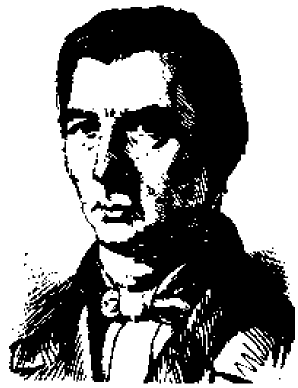
Second, we have noted that an abrupt stiffening of foreign competition *can* hurt U.S. workers by not giving them an adequate chance to adapt gradually to the new conditions. The more rapid the change, the more painful it will be. If it occurs fairly gradually, workers can retrain and move on to the industries that now require their services. If the change is even more gradual, people who retire or leave the threatened industry for other reasons simply need not be replaced. But competition that inflicts its damage overnight is certain to impose real costs upon the affected workers, costs that are no less painful for being temporary. That is why our trade laws make provisions for people and industries damaged by import surges.

UNFAIR FOREIGN COMPETITION

Satire and ridicule are often more persuasive than logic and statistics. Exasperated by the spread of protectionism under the prevailing Mercantilist philosophy, French economist Frédéric Bastiat decided to take the protectionist argument to its illogical conclusion. The fictitious petition of the French candlemakers to the Chamber of Deputies, written in 1845 and excerpted below, has become a classic in the battle for free trade.

> We are subject to the intolerable competition of a foreign rival, who enjoys, it would seem, such superior facilities for the production of light, that he is enabled to inundate our national market at so exceedingly reduced a price, that, the moment he makes his appearance, he draws off all custom for us; and thus an important branch of French industry, with all its innumerable ramifications, is suddenly reduced to a state of complete stagnation. This rival is no other than the sun.
>
> Our petition is, that it would please your honorable body to pass a law whereby shall be directed the shutting up of all windows, dormers, skylights, shutters, curtains, in a word, all openings, holes, chinks, and fissures through which the light of the sun is used to penetrate our dwellings, to the prejudice of the profitable manufactures which we flatter ourselves we have been enabled to bestow upon the country. . . .
>
> We foresee your objections, gentlemen; but there is not one that you can oppose to us . . . which is not equally opposed to your own practice and the principle which guides your policy. . . . Labor and nature concur in different proportions, according to country and climate, in every article of production. . . . If a Lisbon orange can be sold at half the price of a Parisian one, it is because a natural and gratuitous heat does for the one what the other only obtains from an artificial and consequently expensive one. . . .
>
> Does it not argue the greatest inconsistency to check as you do the importation of coal, iron, cheese, and goods of foreign manufacture, merely because and even in proportion as their price approaches *zero*, while at the same time you freely admit, and without limitation, the light of the sun, whose price is during the whole day at *zero*?

SOURCE: F. Bastiat, *Economic Sophisms* (New York: G. P. Putnam's Sons, 1922).



But these are, after all, only qualifications to an overwhelming argument. They call for intelligent monetary and fiscal policies and for transitional assistance to unemployed workers, not for abandonment of free trade. In general, the nation as a whole need not fear competition from cheap foreign labor.

In the long run, labor will be "cheap" only where it is not very productive. Wages will be high in countries with high labor productivity—which holds down costs and permits exporters to compete effectively despite high wages. It is thus misleading to say that the United States held its own in the international marketplace until recently *despite* high wages. Rather it is much more illuminating to point out that the high wages of American workers were a result of high worker productivity, which gave the United States a heavy competitive edge.

We note that in this matter it is *absolute* advantage, not *comparative* advantage, that counts. The country that is most efficient in every output can pay its workers more in every industry.

SUMMARY

1. Countries trade because differences in their natural resources and other inputs create discrepancies in the efficiency with which they can produce different goods, and because **specialization** may offer them greater economies of large-scale production.
2. Voluntary trade will generally be advantageous to both parties in an exchange. This is one of our **Ideas for Beyond the Final Exam.**

3. International trade is more complicated than trade within a nation because of political factors, different national currencies, and impediments to the movement of labor and capital across national borders.
4. Two countries will gain from trade with one another if each exports goods in which it has a **comparative advantage.** Even a country that is generally inefficient will benefit by exporting the goods in whose production it is *least inefficient*. This is another of the **Ideas for Beyond the Final Exam.**
5. When countries specialize and trade, each can enjoy consumption possibilities that exceed its production possibilities.
6. The prices of goods traded between countries are determined by supply and demand, but one must consider explicitly the demand curve and the supply curve of *each* country involved. Thus, in international trade, the equilibrium price must be where the excess of the exporter's quantity supplied over its domestic quantity demanded is equal to the excess of the importer's quantity demanded over its quantity supplied.
7. The *"cheap foreign labor" argument* ignores the principle of comparative advantage, which shows that real wages can rise in both the importing and exporting countries as a result of specialization.
8. **Tariffs** and **quotas** are designed to protect a country's industries from foreign competition. Such protection may sometimes be advantageous to that country, but not if foreign countries adopt tariffs and quotas of their own as a means of retaliation.
9. While the same restriction of trade can be accomplished by either a tariff or a quota, tariffs offer at least two advantages to the country that imposes them: (1) Some of the gains go to the government rather than to foreign producers, and (2) there is greater incentive for efficient production.
10. When a nation shifts from protection to free trade, some industries and their workers will lose out. Equity then demands that these people and firms be compensated in some way. The U.S. government offers protection from import surges and various forms of **trade adjustment assistance** to do this.
11. Several arguments for protectionism can, under the right circumstances, have validity. These include the national defense argument, the *infant-industry argument*, and the use of trade restrictions for *strategic* purposes. But each of these arguments is frequently abused.
12. **Dumping** will hurt certain domestic producers; but it always benefits domestic consumers.

KEY TERMS

Imports
Exports
Specialization
Mutual gains from trade
Absolute advantage
Comparative advantage
"Cheap foreign labor" argument
Tariff
Quota
Export subsidy
Trade adjustment assistance
Infant-industry argument
Strategic trade protection
Dumping

QUESTIONS FOR REVIEW

1. You have a dozen shirts and your roommate has six pairs of shoes worth roughly the same amount of money. You decide to swap six shirts for three pairs of shoes. In financial terms, neither of you gains anything. Explain why you are nevertheless both likely to be better off.
2. In the 18th century, some writers argued that one person in a trade could be made better off only by gaining at the expense of the other. Explain the fallacy in the argument.
3. Country A has a cold climate with a short growing season, but a highly skilled labor force. What sorts of products do you think it is likely to produce? What are the characteristics of the countries with which you would expect it to trade?
4. Upon removal of a quota on sugar, many U.S. sugar farms go bankrupt. Discuss the pros and cons of removing the quota in the short and long runs.
5. Country A's government believes that it is always best to export more than it imports. As a consequence, it exports more to Country B every year than it imports from Country B. After 100 years of this arrangement, both countries are destroyed in an earthquake. What were the advantages and disadvantages of the surplus to Country A? To Country B?
6. The following table describes the number of yards of cloth and barrels of wine that can be produced with a week's worth of labor in England and Portugal. Assume that no other inputs are needed.

	In England	In Portugal
Cloth (yards)	8	12
Wine (barrels)	2	6

a. If there is no trade, what is the price of wine in terms of cloth in England?
b. If there is no trade, what is the price of wine relative to cloth in Portugal?
c. Suppose each country has 1 million weeks of labor available per year. Draw the production possibilities frontier for each country.
d. Which country has an absolute advantage in the production of which good(s)? Which country has a comparative advantage in the production of which good(s)?
e. If the countries start trading with each other, which country will specialize and export which good?
f. What can be said about the price at which trade will take place?

7. Suppose that the United States and Mexico are the only two countries in the world, and that labor is the only productive input. In the United States, a worker can produce 12 bushels of wheat *or* 2 barrels of oil in a day. In Mexico, a worker can produce 2 bushels of wheat *or* 4 barrels of oil per day.
   a. What will be the price ratio between the two commodities (that is, the price of oil in terms of wheat) in each country if there is no trade?
   b. If free trade is allowed and there are no transportation costs, what commodity would the United States import? What about Mexico?
   c. In what range would the price ratio have to fall under free trade? Why?
   d. Picking one possible post-trade price ratio, show clearly how it is possible for both countries to benefit from free trade.

8. The table below presents the demand and supply curves for microcomputers in Japan and the United States.

Price per Computer (thousands of dollars)	Quantity Demanded in United States (thousands)	Quantity Supplied in United States (thousands)	Quantity Demanded in Japan (thousands)	Quantity Supplied in Japan (thousands)
1	90	30	50	50
2	80	35	40	55
3	70	40	30	60
4	60	45	20	65
5	50	50	10	70
6	40	55	0	75

   a. Draw the demand and supply curves for the United States on one diagram and those for Japan on another one.
   b. If there is no trade between the United States and Japan, what are the equilibrium price and quantity in the computer market in the United States? In Japan?
   c. Now suppose trade is opened up between the two countries. What will be the equilibrium price in the world market for computers? What has happened to the price of computers in the United States? In Japan?
   d. Which country will export computers? How many?
   e. When trade opens, what happens to the quantity of computers produced, and therefore employment, in the computer industry in the United States? In Japan? Who benefits and who loses *initially* from free trade?

9. Under current trade law, the president of the United States must report periodically to Congress on countries engaging in unfair trade practices that inhibit U.S. exports. How would you define an "unfair" trade practice? Suppose Country X exports much more to the United States than it imports, year after year. Does that constitute evidence that Country X's trade practices are unfair? What would constitute such evidence?

10. Suppose the United States finds Country X guilty of unfair trade practices and penalizes it with import quotas. So U.S. imports from Country X fall. Suppose, further, that Country X does not alter its trade practices in any way. Is the United States better or worse off? What about Country X?



CHAPTER 35

THE INTERNATIONAL MONETARY SYSTEM: ORDER OR DISORDER?

Cecily, you will read your Political Economy in my absence. The chapter on the Fall of the Rupee you may omit. It is somewhat too sensational.
Miss Prism in Oscar Wilde, *The Importance of Being Earnest*

The last chapter discussed the reasons for international trade and the benefits that accrue to all nations when countries specialize in producing goods in which they have a comparative advantage. But when goods cross national borders, *money* must generally move in the opposite direction. When the United States buys coffee from Brazil, we must send money to the Brazilians. When Japan purchases petroleum from Saudi Arabia, it must send money to the Saudis, and so on. This chapter is about the system that has been set up to handle these international movements of money—the *international monetary system.*

We begin by investigating a system in which rates of exchange among national currencies are determined in free markets by the laws of supply and demand. We shall see that the main macroeconomic variables that we have been studying—output, the price level, and the rate of interest—each play a role in the determination of a country's exchange rate. This discussion sets the stage for Chapter 36, where we will learn how movements of the exchange rate, in turn, affect the national economy.

Next, we turn to the opposite polar form—an international monetary system in which exchange rates are fixed by government authority, rather than by the market. The United States does not now fix its exchange rate, but other countries do. And studying such a system will help us understand why some people believe that the world should move toward greater fixity in exchange rates.

PUZZLE: WHATEVER HAPPENED TO MEXICO?

For much of 1994, Mexico was considered an economic success story: a model of how a developing country can turn its economy around by pursuing sound economic policies. The Mexican government had balanced its budget, opened its economy to world trade, and brought inflation down from over 100 percent per year to under 5 percent. Furthermore, it was guaranteeing the soundness of its currency by pledging to keep the value of the peso fixed in terms of U.S. dollars. With

the North American Free Trade Agreement (NAFTA) bringing new opportunities to Mexico's now-vibrant export industries, rapid growth seemed assured.

Then, around Christmas of 1994, the bottom fell out. With a suddenness that was both stunning and vicious, massive amounts of money started flowing out of Mexico. Interest rates soared, and the stock market collapsed. The value of the peso, which had been fixed at about 29 cents, tumbled to 19 cents within a few weeks and to about 14 cents within a few months. Soon inflation, which had been down to about 4 percent, was soaring (for a few months) at annual rates around 100 percent. And Mexico plunged into a severe recession.

What went wrong? What forced the Mexican government to abandon its fixed exchange rate policy? And why did that decision have such dire consequences? We will learn some of the answers in the pages to come. But first we need to understand what determines exchange rates.

WHAT ARE EXCHANGE RATES?

We noted in the previous chapter that international trade is more complicated than domestic trade. There are no national borders to be crossed when, say, California lettuce is shipped to Massachusetts. The consumer in Boston pays with *dollars*, just the currency that the farmer in Modesto wants. But if that same farmer ships his lettuce to Japan, consumers there will have only Japanese *yen* with which to pay, rather than the dollars the farmer in California wants. Thus, if international trade is to take place, there must be a way to transform one currency into another. The rates at which such transformations are made are called **exchange rates.**

The **EXCHANGE RATE** states the price, in terms of one currency, at which another currency can be bought.

There is an exchange rate between every pair of currencies. For example, $1 is currently the equivalent of about 5 French francs. The exchange rate between the franc and the dollar, then, may be expressed as roughly "5 francs to the dollar" (meaning that it costs 5 francs to buy a dollar) or about "20 cents to the franc" (meaning that it costs 20 cents to buy a franc).

There have been some dramatic changes in the international value of the dollar over time. In a nutshell, the dollar soared in the period from mid-1980 to early 1985, fell against most major currencies from early 1985 until early 1988, and has generally fluctuated without any pronounced trend since. This chapter seeks to explain such currency movements.

Under our present system, currency rates change frequently. When other currencies become more expensive in terms of dollars, we say that they have **appreciated** relative to the dollar. Alternatively, we can look at this same event as the dollar buying less foreign currency, meaning that the dollar has **depreciated** relative to another currency.

A nation's currency is said to **APPRECIATE** when exchange rates change so that a unit of its own currency can buy more units of foreign currency.

What is a depreciation to one country must be an appreciation to the other.

The currency is said to **DEPRECIATE** when exchange rates change so that a unit of its currency can buy fewer units of foreign currency.

For example, if the cost of a German mark rises from 50 cents to 60 cents, the cost of a U.S. dollar in terms of marks simultaneously falls from 2.00 marks to 1.67 marks. The Germans have had a currency *appreciation* while we have had a currency *depreciation*.

Notice also that, when many currencies are changing in value, the dollar may be appreciating with respect to one currency but depreciating with respect to another. Table 35-1 gives an indication of exchange rates prevailing in July 1980, February 1985, and April 1996, showing how many dollars or cents it cost at each of those times to buy each unit of foreign currency. You will note that, between

TABLE 35-1 EXCHANGE RATES WITH THE U.S. DOLLAR (DOLLARS PER UNIT OF FOREIGN CURRENCY)

Country	Currency Unit	Symbol	Cost in Dollars July 1980	February 1985	April 1996
Australia	dollar	$	$1.16	$0.74	$0.79
Canada	dollar	$	0.87	0.74	0.73
France	franc	FF	0.25	0.10	0.19
Germany	mark	DM	0.57	0.30	0.66
Italy	lira	L	0.0012	0.00049	
Japan	yen	¥	0.0045	0.0038	0.00064
Mexico	new peso	$	44.0[a]	5.0[a]	0.0094
Sweden	krona	Kr	0.24	0.11	0.14
Switzerland	franc	S.Fr.	0.62	0.36	0.15
United Kingdom	pound	£	2.37	1.10	0.81

[a]On January 1, 1993, the peso was redefined so that 1,000 old pesos were equal to 1 new peso. Hence, the numbers 44 and 5 listed for July 1980 and February 1985 were actually 0.044 and 0.005 on the old basis.

SOURCE: International Financial Statistics and *The Wall Street Journal*.

February 1985 and April 1996, the dollar *depreciated* sharply relative to the Japanese yen and most European currencies. For example, the British pound rose from $1.10 to $1.51. Yet during that same period the dollar *appreciated* dramatically relative to the Mexican peso; it bought about 0.2 pesos in 1985 but over 7.1 in 1996.[1]

A **DEVALUATION** is a reduction in the official value of a currency.

A **REVALUATION** is an increase in the official value of a currency.

While this is the terminology used to describe movements of exchange rates in free markets, another set of terms is used to describe decreases and increases in currency values when those values are set by government decree. When an officially set exchange rate is altered so that a unit of a nation's currency can buy *fewer* units of foreign currency, we say there has been a **devaluation** of that currency. When the exchange rate is altered so that the currency can buy *more* units of foreign currency, we say there has been a **revaluation**.

EXCHANGE RATE DETERMINATION IN A FREE MARKET

FLOATING EXCHANGE RATES are rates determined in free markets by the law of supply and demand.

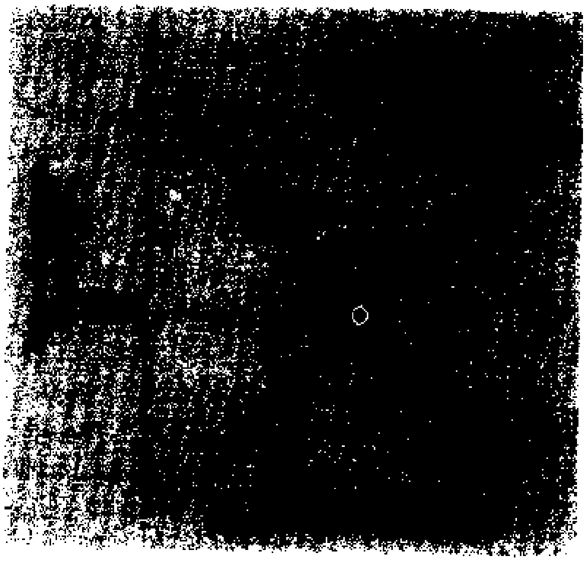
Why is it that a German mark now costs about 66 cents and not 60 cents or 75 cents? In a world of **floating exchange rates**, with no government interferences, the answer would be straightforward. Exchange rates would be determined by the forces of supply and demand, just like the prices of apples, computers, and haircuts.

In a leap of abstraction, imagine that the United States and Germany are the only countries on earth, so there is only one exchange rate to be determined. Figure 35-1 depicts the determination of this exchange rate at the point (denoted *E* in the figure) where demand curve *DD* crosses supply curve *SS*. At this price (70 cents per mark), the number of marks demanded is equal to the number of marks supplied.

[1]In fact, the dollar bought about 200 pesos in February 1985, but that is because the old peso was replaced by a new peso in January 1993, which moved the decimal point three places.

FIGURE 35-1 DETERMINATION OF EXCHANGE RATES IN A FREE MARKET

Like any price, an exchange rate in a free market will be determined by the intersection of the demand and supply curves. Point *E* depicts this point for the exchange rate between the U.S. dollar and the German mark, which settles at 70 cents per mark in this example.



In a free market, exchange rates are determined by the law of supply and demand. If the rate were below the equilibrium level, the quantity of marks demanded would exceed the quantity of marks supplied, and the price of a mark would be bid up. If the rate were above the equilibrium level, quantity supplied would exceed quantity demanded, and the price of a mark would fall. Only at the equilibrium exchange rate is there no tendency for the rate to change.

As usual, supply and demand determine price. But in this case, we must ask: Where do the supply and demand come from? Why does anyone demand a German mark? The answer has three parts:

1. *International trade in goods and services.* This was the subject of the last chapter. If, for example, Jane Doe, an American, wants to buy a German automobile, she will first have to buy marks with which to pay the dealer in Munich.[2] So Jane's demand for a German *car* leads to a demand for German *marks.* In general, *demand for a country's exports leads to a demand for its currency.*[3]
2. *International trade in financial instruments like stocks and bonds.* For example, if American investors want to purchase German stocks, they will first have to acquire the marks that the sellers will insist upon. In this way, demand for German financial assets leads to demand for German marks. Thus, *demand for a country's financial assets leads to a demand for its currency.*
3. *Purchases of physical assets like factories and machinery overseas.* If IBM wants to buy out a small German computer manufacturer, the owners will no doubt want to receive marks. So IBM will first have to acquire German currency. In general, *direct foreign investment leads to a demand for a country's currency.*

---

[2]Actually, she will not do this because banks generally handle foreign exchange transactions for consumers. An American bank probably will buy the marks for her. But the effect is exactly the same as if Jane had done it herself.

[3]See Review Question 2 at the end of the chapter (page 843).

Now, where does the supply come from? To answer this, just turn all of these transactions around. Germans wanting to buy U.S. goods and services, or invest in U.S. financial markets, or make direct investments in America will have to offer their marks for sale in the foreign-exchange market (which is similar to the stock market) to acquire the needed dollars. To summarize:

The *demand* for a country's currency is derived from the demands of foreigners for its export goods and services and for its assets—including financial assets, like stocks and bonds, and real assets, like factories and machinery. The *supply* of a country's currency arises from its imports, and from foreign investment by its own citizens.

To appreciate the usefulness of even this simple supply and demand analysis, think about how the exchange rate between the dollar and the mark would change if there were an economic boom in the United States. One important effect of such a boom would be to stimulate American demand for German products, such as machine tools, cameras, and beer. In terms of the supply-demand diagram shown in Figure 35-2, the increased desires of Americans for German goods would shift the demand curve for German marks out from D_1D_1 (the black line in the figure) to D_2D_2 (the blue line). Equilibrium would shift from point E to point A, and the exchange rate would rise from 70 cents per mark to 75 cents per mark. The increased demand for marks by U.S. citizens causes the mark to *appreciate* relative to the dollar.

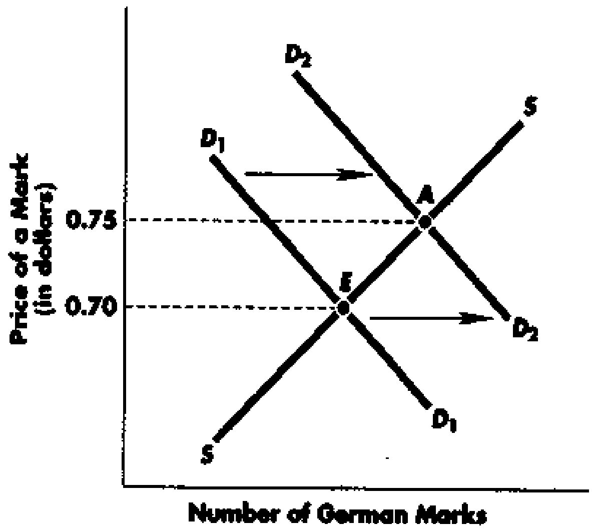
EXERCISE:

Test your understanding of the supply and demand analysis of exchange rates by showing why each of the following events would lead to an appreciation of the mark (a depreciation of the dollar) in a free market:

1. A recession in Germany cuts German purchases of American goods.
2. American investors are attracted by prospects for profit on the German stock market.

FIGURE 35-2 THE EFFECT OF AN ECONOMIC BOOM ON THE EXCHANGE RATE

If the U.S. economy suddenly booms, Americans will spend more on imports from Germany. Thus, the demand curve for German marks will rise from D_1D_1 to D_2D_2 as Americans seek to acquire the marks they need. The diagram shows that this will cause the mark to appreciate, from 70 cents to 75 cents, as equilibrium shifts from point E to point A. Looked at from the U.S. perspective, the dollar will depreciate.



3. Interest rates on government bonds rise in Germany but are stable in the United States. (*Hint:* Which country's citizens will be attracted by high interest rates in the other country?)

To say that supply and demand determine exchange rates in a free market is at once to say everything and to say nothing. If we are to understand the reasons why some currencies appreciate while others depreciate, we must look into the factors that move the supply and demand curves. Economists believe that the principal determinants of exchange rate movements are rather different in the long, medium, and short runs. So we turn in the next three sections to the analysis of exchange rate movements over these three "runs," beginning with the long run.

THE PURCHASING-POWER PARITY THEORY: THE LONG RUN

As long as there is free trade across national borders, exchange rates should eventually adjust so that the same product costs the same whether measured in dollars in the United States, marks in Germany, yen in Japan, and so on—except for differences in transportation costs and the like. This simple statement forms the basis of the major theory of exchange rate determination in the long run.

The *purchasing-power parity theory of exchange rate determination* holds that the exchange rate between any two national currencies adjusts to reflect differences in the price levels in the two countries.

An example will bring out the basic truth in this theory and also suggest some of its limitations. Suppose that German and American steel are identical and that these two nations are the only producers of steel for the world market. Suppose further that steel is the only tradable good that either country produces.

Question: If American steel costs $210 per ton and German steel costs 300 marks per ton, what must be the exchange rate between the dollar and the mark?

Answer: Since 300 marks or $210 each buys a ton of steel, they must be of equal value. Hence, each mark must be worth 70 cents. Why? Any higher price for a mark, like 80 cents, would mean that steel would cost $240 per ton (300 marks at 80 cents each) in Germany but only $210 per ton in the United States. In that case, all foreign customers would shop for their steel in the United States—which would increase the demand for dollars and decrease the demand for marks. Similarly, any exchange rate below 70 cents per mark would send all the steel business to Germany, driving the value of the mark up toward its purchasing-power parity level.

EXERCISE:
Show why an exchange rate of 60 cents per mark is too low.

The purchasing-power parity theory is used to make long-run predictions about the effects of inflation on exchange rates. To continue our example, suppose that over a 10-year period, prices in the United States rise by 50 percent while prices in Germany rise by 110 percent. The purchasing-power parity theory predicts that the mark will depreciate relative to the dollar. It also predicts the amount of the depreciation. After the inflation, American steel costs $315 per ton (50 percent more than $210), while German steel costs 630 marks per ton (110 percent more than 300 marks). For these two prices to be equivalent, 630 marks

must be worth $315, or 1 mark must be worth 50 cents. The mark, therefore, must have fallen from 70 cents to 50 cents.

According to the purchasing-power parity theory, differences in domestic inflation rates are a major cause of adjustments in exchange rates. If one country has higher inflation than another, then its exchange rate should be depreciating.

For many years, the theory seemed to work tolerably well. While precise numerical predictions based on purchasing-power parity calculations were never very accurate (see the box on the next page), nations with higher inflation did at least experience depreciating currencies. But in the 1980s, even this broke down. For example, while the U.S. inflation rate was higher than both Germany's and Japan's, the dollar nonetheless rose sharply relative to both the mark and the yen from 1980 to 1985. Clearly, the theory was missing something. What?

Many things. But perhaps the principal failing of the purchasing-power parity theory is that it focuses too much on trade in goods and services. Financial assets like stocks and bonds are also traded actively across national borders—and in vastly greater dollar volumes than trade in goods and services. If investors decide that, say, U.S. assets are a better bet than German assets, the dollar will rise even if our inflation rate is well above Germany's. For this and other reasons:

Most economists believe that other factors are much more important than relative price levels for exchange rate determination in the short run. But in the long run, purchasing-power parity plays an important role.

Economic Activity and Exchange Rates: The Medium Run

Since consumer spending increases when income rises and decreases when income falls, the same is likely to happen to spending on imported goods. For this reason:

A country's imports will rise quickly when its economy is booming and slowly when its economy is stagnating.

We have already illustrated this point with Figure 35-2. There we saw that a boom in the United States would shift the demand curve for marks outward as Americans bought more German goods. And that, in turn, would lead to an appreciation of the mark (depreciation of the dollar) as Americans sold dollars to buy marks. However, if Germany were booming at the same time, German citizens would be buying more American exports, which would shift the supply curve of marks outward. On balance, the value of the dollar might or might not fall. What matters is whether exports are growing faster than imports. The general lesson is that:

Holding other things equal, a country whose aggregate demand grows faster than the rest of the world's normally finds its imports growing faster than its exports. Thus, its demand curve for foreign currency shifts outward more rapidly than its supply curve, causing its currency to depreciate.

This is one reason why it is unwise to interpret a "strong currency" as an indication of a "strong economy." A nation that grows more rapidly than its trading partners may find itself with a depreciating currency.

PURCHASING-POWER PARITY AND THE BIG MAC

Since 1986, *The Economist* magazine has been using a well-known international commodity—the Big Mac—to assess the purchasing-power parity theory of exchange rates, or as the magazine puts it, "to make exchange-rate theory more digestible." The famous burger is now available in 79 countries. The 1995 survey included 32 of them, and it showed that the theory does not always work terribly well.

Here's how the survey works. In the spring of 1995, the average price of a Big Mac in the United States was $2.32, including sales tax. In Japan, the same commodity cost ¥391. For those two amounts to be equal, a dollar would have had to be worth about ¥391/$2.32 = ¥169. But the actual value of the dollar at the time was just ¥84, meaning that the Big Mac indicates an overvalued yen relative to the dollar. (In fact, the dollar rose about 28 percent against the yen in the year following the survey.)

The same calculation using the British pound yields much the same conclusion. By the Big Mac standard, the pound was overvalued by 21 percent. At £1.74 per burger, the implied purchasing-power parity value of the pound was $2.32/£1.74 = $1.33, versus a market value of $1.61.

True Big Mac aficionados will find the following data helpful, especially if they are planning international travel. The cheapest Big Macs can be found in China; the 9 yuan price was a mouthwatering dollar equivalent of $1.05. The Hong Kong version is almost as good, weighing in at just $1.23. The priciest Big Macs in the world were to be found in Switzerland and Denmark, at the local equivalents of $5.20 and $4.92, respectively.

SOURCE: "Big MacCurrencies," *The Economist*, April 15, 1995, p. 74.



THE HAMBURGER STANDARD

	Big Mac Prices				
	In Local Currency	In Dollars	Implied PPP[a] of the Dollar	Actual $ Exchange Rate 4/7/95	Local Currency Under (−)/Over (+) Valuation, %
United States[b]	$2.32	2.32	—	—	—
Britain	£1.74	2.80	1.33[c]	1.61[c]	+21
China	Yuan9.00	1.05	3.88	8.54	−55
Denmark	DKr26.75	4.92	11.5	5.43	+112
France	FF18.5	3.85	7.97	4.80	+66
Hong Kong	HK$9.50	1.23	4.09	7.73	−47
Italy	Lire4,500	2.64	1,940	1,702	+14
Japan	¥391	4.65	169	84.2	+100
Mexico	Peso10.9	1.71	4.70	6.37	−26
Switzerland	S.Fr.5.90	5.20	2.54	1.13	+124

[a]Purchasing-power parity: local price divided by price in the United States.

[b]Average of New York, Chicago, San Francisco, and Atlanta.

[c]Dollars per pound.

SOURCE: McDonald's; reported in "Big MacCurrencies," *The Economist* April 15, 1995, p. 74.

INTEREST RATES AND EXCHANGE RATES: THE SHORT RUN

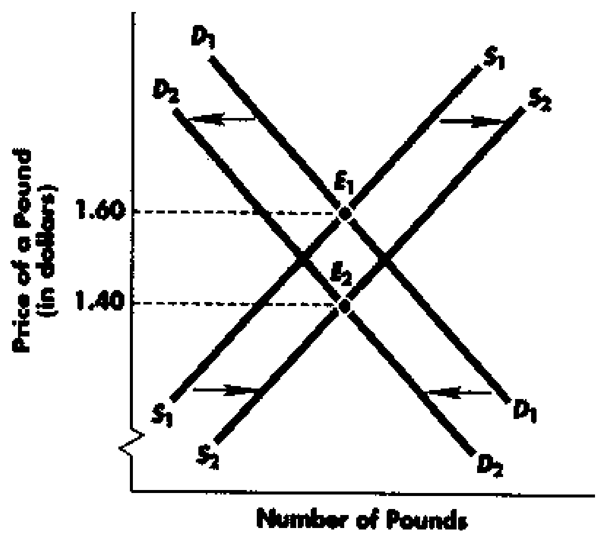
While economic activity is important for exchange rate determination in the medium run, "other things" often are not equal in the short run. Specifically, one factor that often seems to call the tune in determining exchange rate movements in the short run is *interest rate differentials*. A multi-trillion dollar pool of so-called *hot money*—owned by banks, mutual funds, multinational corporations, and wealthy individuals of all nations—travels rapidly around the globe in search of the highest interest rates.

Thus, suppose that British government bonds are paying a 6 percent rate of interest when yields on equally safe American government securities rise to 8 percent. British investors will be attracted by the high interest rates in the United States and will offer pounds for sale in order to buy dollars, planning to use those dollars to buy American securities. At the same time, American investors will find investing in the United States more attractive than ever, so fewer pounds will be demanded by Americans.

When the demand schedule falls and the supply curve rises, the effect on price is predictable: The pound will depreciate, as Figure 35-3 shows. In the figure, the supply curve of pounds shifts outward from S_1S_1 to S_2S_2 when British investors seek to sell pounds in order to purchase U.S. securities. At the same time, American investors wish to buy fewer pounds because they no longer wish to invest in British securities. Thus, the demand curve shifts inward from D_1D_1 to D_2D_2. The result, in our example, is a depreciation of the pound from \$1.60 to \$1.40. In general:

Holding other things equal, countries that offer investors higher rates of return are able to attract more capital than are countries that offer lower rates. Thus, a rise in interest rates often will lead to an appreciation of the currency, and a drop in interest rates will lead to a depreciation.

FIGURE 35-3 THE EFFECT OF A RISE IN U.S. INTEREST RATES



When the United States raises its interest rates, more British investors will want to buy American bonds, and so the supply curve of pounds will shift outward from S_1S_1 to S_2S_2. At the same time, fewer Americans will seek to buy British bonds, so the demand curve for pounds will shift inward from D_1D_1 to D_2D_2. The combined effect of these two shifts is to move the market equilibrium from point E_1 to point E_2. The British pound depreciates, and the dollar appreciates.

Most experts in international finance agree that this factor is the major determinant of exchange rates in the short run. It has certainly played a predominant role in the stunning movements of the U.S. dollar since the 1980s. In the early 1980s, American interest rates rose well above comparable interest rates abroad. In consequence, foreign capital was attracted here, American capital stayed at home, and the dollar soared. Something quite similar happened to the German mark in the early 1990s.

MARKET DETERMINATION OF EXCHANGE RATES: SUMMARY

We can summarize this discussion of exchange rate determination in free markets as follows:

1. Currency values generally will be *appreciating* in countries whose inflation rates are lower than the rest of the world's because buyers in foreign countries will demand their goods and thus drive up their currencies.
2. Exchange rates would also be expected to rise in countries where aggregate demand is growing more slowly than average, because these countries will be importing rather little.
3. We expect to find appreciating currencies in countries which offer high rates of return to investors because these countries will attract capital from all over the world.

Reversing each of these, we expect that currencies will be *depreciating* in countries with relatively high inflation rates, or rapid demand growth, or low interest rates.

FIXED EXCHANGE RATES AND THE BALANCE OF PAYMENTS

Some exchange rates today are truly floating, determined by the forces of supply and demand without government interference. But many are not. Furthermore, some people claim that exchange rate fluctuations are so troublesome that the world would be better off with fixed exchange rates. For these reasons, we turn our attention next to a system of **fixed exchange rates**, or rates that are set by government. Naturally, under such a system the exchange rate, being fixed, is not closely watched. Instead, international financial specialists focus on a country's *balance of payments*—a term we must now define.

FIXED EXCHANGE RATES are rates set by government decisions and maintained by government actions.

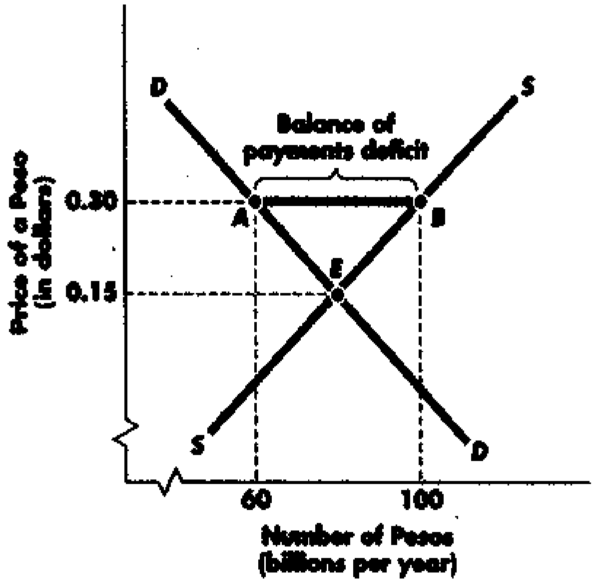
To understand what the balance of payments is, look at Figure 35-4, which depicts a situation that might represent, say, Mexico just before the peso tumbled in the winter of 1994 to 1995—an *overvalued* currency. While the supply and demand curves for pesos indicate an equilibrium exchange rate of 15 cents to the peso (point *E*), the Mexican government is keeping the rate at 30 cents. Notice that, at 30 cents per peso, more people are supplying pesos than are demanding them. In the example, suppliers are selling 100 billion pesos per year, but purchasers want only 60 billion.

The **BALANCE OF PAYMENTS DEFICIT** is the amount by which the quantity supplied of a country's currency (per year) exceeds the quantity demanded. Balance of payments deficits arise whenever the exchange rate is pegged at an artificially high level.

This gap between the 100 billion pesos that some people sell and the 60 billion that others buy is what we mean by Mexico's **balance of payments deficit**—40 billion pesos per year in this case. It is shown by the horizontal distance between points *A* and *B* in Figure 35-4.

FIGURE 35-4 A BALANCE OF PAYMENTS DEFICIT

At a fixed exchange rate of 30 cents per peso, which is well above the equilibrium level of 15 cents per peso, Mexico's currency is overvalued in this example. As a consequence, more pesos will be supplied (point *B*) than are demanded (point *A*). The difference—distance *AB*, or 40 billion pesos per year—represents Mexico's balance of payments deficit.



How can market forces be flouted in this way? Since sales and purchases on any market must be equal, as a simple piece of arithmetic, the excess of quantity supplied over quantity demanded of Mexican currency (40 billion pesos per year in this example) must be bought by the Mexican government. To buy up these pesos, it must give up some of the foreign currency that it holds as *reserves*. Thus, the Bank of Mexico would be losing about $12 billion in reserves per year as the cost of keeping the peso at 30 cents.

Naturally, this cannot go on forever; the reserves eventually will run out. And this is the fatal flaw in a system of fixed exchange rates. Once speculators become convinced that the exchange rate can be held for only a short while longer, they will sell the overvalued currency in massive amounts rather than hold on to money whose value they expect to fall.

That is precisely what happened to Mexico in the winter of 1994 to 1995. The supply curve of pesos shifted outward drastically, as shown in Figure 35-5, causing a sharp rise in the balance of payments deficit—from 40 billion to 90 billion pesos in the example. (The numbers are fictitious.) Lacking sufficient reserves, the Mexican government had no choice but to let the peso fall to its equilibrium level, and this probably amounted to an even larger devaluation than would have been required before the speculative "run" on the peso.

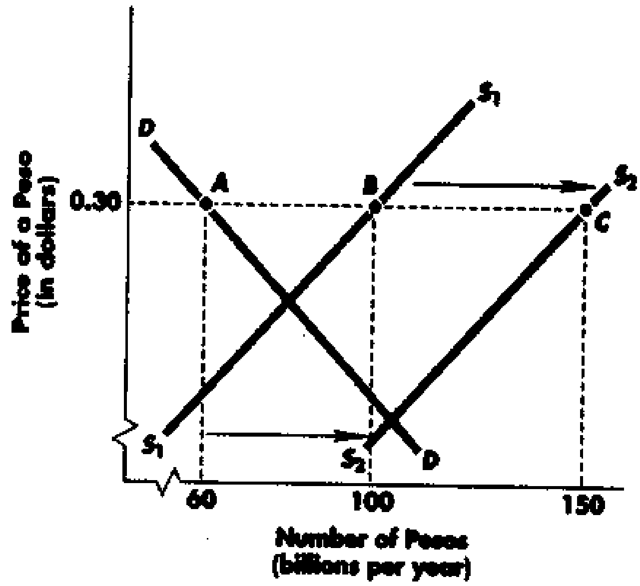
The **BALANCE OF PAYMENTS SURPLUS** is the amount by which the quantity demanded of a country's currency (per year) exceeds the quantity supplied. Balance of payments surpluses arise whenever the exchange rate is pegged at an artificially low level.

For an example of the reverse case, a severely *undervalued* currency, let us go back in history to the case of West Germany in 1973. Figure 35-6 depicts demand and supply curves for marks that intersect at an equilibrium price of 50 cents per mark (point *E* in the diagram). Yet, in the example, we suppose that the German authorities are holding the rate at 33 cents. At this rate, the quantity of marks demanded (50 billion) greatly exceeds the quantity supplied (40 billion). The difference is Germany's **balance of payments surplus**, and is shown by the horizontal distance *AB*.

Germany can keep the rate at 33 cents only by providing the marks that foreigners want to buy: 10 billion marks per year in this example. In return, it receives U.S. dollars, British pounds, French francs, gold, and so on. All of this

FIGURE 35-5 A SPECULATIVE RUN ON THE PESO

When speculators become convinced that a devaluation of the peso is in the offing, they will rush to sell pesos. Their actions shift the supply curve outward from S_1S_1 to S_2S_2 and, in the process, widen the Mexican balance of payments deficit from *AB* to *AC*.



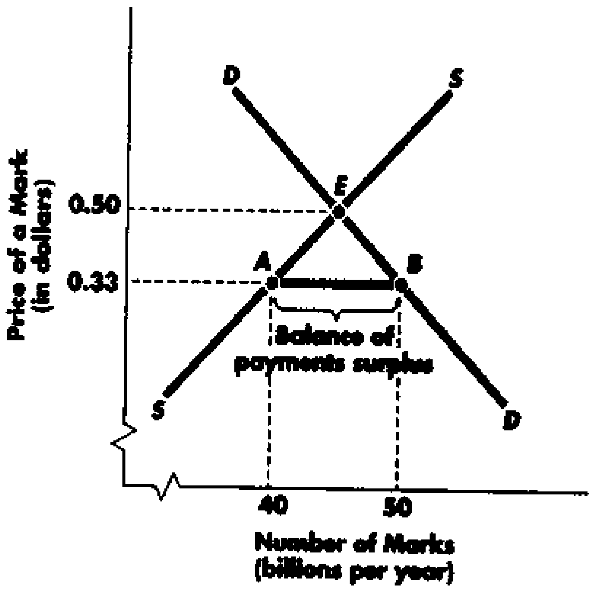
serves to increase Germany's reserves of foreign currencies. But notice the important difference between this case and the overvalued U.S. dollar.

The accumulation of reserves rarely will *force* a central bank to revalue in the way that depletion of reserves can force a devaluation.

This asymmetry was a clear weakness of the system of fixed exchange rates that prevailed between 1944 and 1971. In principle, an imbalance in exchange rates could be cured either by a *devaluation* by the country with a balance of payments deficit or by an upward *revaluation* by the country with a balance of

FIGURE 35-6 A BALANCE OF PAYMENTS SURPLUS

In this example, Germany's currency is undervalued at 33 cents per mark since the equilibrium exchange rate is 50 cents per mark. Consequently, more marks are being demanded (point *B*) than are being supplied (point *A*). The gap between quantity demanded and quantity supplied—distance *AB*, or 10 billion marks per year—measures Germany's balance of payments surplus.



payments surplus. In practice, though, it was almost always the deficit countries that were forced to act.

Why did the surplus countries refuse to revalue? One reason was a stubborn refusal to recognize some basic economic realities. They viewed the disequilibrium as the problem of the deficit countries and believed that the deficit countries, therefore, should take the corrective steps. This, of course, is nonsense in a worldwide system of fixed exchange rates. Some currencies are overvalued *because* some other currencies are undervalued. In fact, the two statements mean exactly the same thing.

The other reason why exporters in Germany, Japan, and other surplus countries resisted upward revaluations is that such actions would have made their products more expensive to foreigners and thus cut into their sales. And these exporters had the political clout to make their views stick. Meanwhile, since the values of the mark and the yen on world markets were artificially held down, German and Japanese consumers were put in the unenviable position of having to pay more for imported goods than they need have paid. Rather than buy these excessively expensive foreign goods, they watched domestically produced goods go overseas in return for pieces of paper (dollars, francs, pounds, and so on).

DEFINING THE BALANCE OF PAYMENTS IN PRACTICE

The preceding discussion makes it look simple to measure a nation's balance of payments position: You just count up the private demand for and supply of its currency and subtract quantity supplied from quantity demanded. Conceptually, that is all there is to it. But in practice the difficulties are great because we never observe directly the number of dollars demanded and supplied.

If we look at actual market transactions, we will see that the number of, say, U.S. dollars actually *purchased* and the number of U.S. dollars actually *sold* are identical. Unless someone has made a bookkeeping error, this must always be so. How, then, can we recognize a balance of payments surplus or deficit? Easy, you say. Just look at the transactions of the central bank, whose purchases or sales must make up the difference between private demand and private supply. If the Federal Reserve is buying dollars, its purchases measure our balance of payments deficit. If it is selling, its sales represent our balance of payments surplus.

Thus, the suggestion is to measure the balance of payments by *excluding official transactions among governments*. This is roughly how the balance of payments surplus or deficit is defined today although, for a variety of complicated reasons, the U.S. government decided long ago to stop publishing any official statistic called "the balance of payments deficit." Instead, it lists all foreign transactions and invites readers to define the balance of payments any way they wish. Many experts focus on the surplus or deficit in what is called the *current account*. This account includes all exports and imports of goods and services, all cross-border payments of interest and dividends, and all cross-border gifts from and to individuals or governments. For more than a decade now, the United States has run large current account deficits.

But this hardly represents our "balance of payments," as it leaves out all purchases and sales of assets. Purchases of U.S. assets by foreigners bring us foreign currency, and U.S. purchases of foreign assets cost us foreign currency. Netting the capital flows in each direction gives us our surplus or deficit on *capital account*. In recent years, this element of our balance of payments has registered

persistently large surpluses, as the United States borrowed large amounts of money from abroad.

The balance of payments can balance in one of two ways. If the government does not intervene to fix the exchange rate, all private transactions—current account plus capital account—must add up to zero (dollars purchased = dollars sold). But if, instead, the exchange rate is fixed, as shown in Figures 35-4 and 35-6, the two accounts need not balance one another. Government purchases or sales of foreign currency make up the surplus or deficit in the overall balance of payments.

A BIT OF HISTORY: THE GOLD STANDARD

It is hard to find examples of strictly fixed exchange rates in the historical record. About the only time exchange rates were truly fixed was under the old *gold standard*, at least when it was practiced in its ideal form.[4]

Under the gold standard, fixed exchange rates were maintained by an automatic equilibrating mechanism that went something like this: All currencies were defined in terms of gold; indeed, some were actually made of gold. When a nation had a deficit in its balance of payments, this meant, essentially, that more gold was flowing *out* than was flowing *in*. Since the domestic money supply was based on gold, losing gold to foreigners meant that the quantity of money fell *automatically*. This raised interest rates and attracted foreign capital. At the same time, the restrictive "monetary policy" pulled down output and prices, thus discouraging imports and encouraging exports. The balance of payments problem quickly rectified itself. This meant, however, that:

Under the gold standard, no nation had control of its domestic monetary policy, and therefore no nation could control its domestic economy very well.

At least in principle, the effects on surplus countries were perfectly symmetrical under the gold standard. A balance of payments surplus led to gold inflows, and thus to an increase in the domestic money supply, whether the surplus country liked the idea or not. This raised prices and output, thereby increasing imports and decreasing exports. And it also lowered interest rates, thereby encouraging outflows of capital. Because of these automatic adjustments, nations rarely reached the point at which devaluations or revaluations were necessary. Exchange rates were fixed as long as countries abided by the rules of the gold standard game.

In addition to the loss of control over domestic monetary conditions, the gold standard posed one other serious difficulty.

A fundamental problem with the gold standard was that the world's commerce was at the mercy of gold discoveries.

Discoveries of gold meant higher prices in the long run and higher real economic activity in the short run, through the standard monetary-policy mechanisms that we studied in the previous part of this book. And when the supply of gold did not keep pace with growth of the world economy, prices had to fall in the long run and employment had to fall in the short run.

---

[4]As a matter of fact, while the gold standard lasted (on and off) for hundreds of years, it was rarely practiced in its ideal form. Except for a brief period of fixed exchange rates in the late 19th and early 20th centuries, there were periodic adjustments of exchange rates even under the gold standard.

THE BRETTON WOODS SYSTEM AND THE INTERNATIONAL MONETARY FUND

The gold standard faltered many times and finally collapsed amid the financial chaos of the Great Depression of the 1930s and World War II. Without it, the world struggled through a serious breakdown in international trade.

Then, as World War II drew to a close, with much of Europe in ruins and with the United States holding the lion's share of the free world's reserves, officials of the industrial nations met at Bretton Woods, New Hampshire, in 1944. Their goal was to establish a stable monetary environment that would facilitate world trade. Since the dollar was the only "strong" currency at that time, it was natural to turn to the dollar as the basis of the new international economic order. And that is just what they did.

The Bretton Woods agreements reestablished a system of fixed exchange rates based not on the old gold standard but on the free convertibility of the U.S. dollar into gold. The United States agreed to buy or sell gold to maintain the $35 per ounce price that had been established by President Franklin Roosevelt in 1933. The other signatory nations, which had almost no gold in any case, agreed to buy and sell *dollars* to maintain their exchange rates at agreed-upon levels. Thus, all currencies were indirectly on a modified "gold standard." A holder of French francs, for example, could exchange these for dollars at (roughly) 5 francs per dollar and then exchange these into gold at $35 per ounce. In this way, the value of the franc was fixed at 175 francs per ounce of gold (5 francs per dollar times 35 dollars per ounce). The new system was dubbed the *Bretton Woods system.*

The *International Monetary Fund (IMF)* was set up to police and manage this new system. Using funds that had been contributed by member countries, the IMF was empowered to make loans to countries that were running low on reserves. A change in exchange rates was to be permitted only in the case of a "fundamental disequilibrium" in a nation's balance of payments—for it was believed that only relatively fixed exchange rates could provide the stable climate needed to restore world trade.

Of course, the Bretton Woods conferees did not define clearly what a "fundamental disequilibrium" was, nor could they have. As the system evolved, it came to mean a chronic *deficit* in the balance of payments of sizable proportions. Such nations would then *devalue* their currencies relative to the dollar. So the system was not really one of fixed exchange rates but rather one where rates were "fixed until further notice."

The Bretton Woods system had many of the flaws that we mentioned in our discussion of the pure system of fixed exchange rates. First, since devaluations were permitted only after a long run of balance of payments deficits, these devaluations (a) could be clearly foreseen and (b) normally had to be large. Speculators then saw opportunities for profit and would "attack" weak currencies with waves of selling.

This problem led many economists to question whether the system of fixed exchange rates was really providing the stable climate for world trade that had been intended. Was a system where rates were constant for long periods and then altered by large amounts really more conducive to international trade than one where overvalued currencies would gradually depreciate, as they would under a system of floating rates?

The second problem arose from the custom that deficit nations were expected to devalue when forced to, while surplus nations could resist upward revaluations. Since the U.S. dollar defined the monetary value of gold (at $35 per ounce),

America was the one nation in the world that had no way to devalue its currency relative to gold, no matter how "fundamental" the disequilibrium became. The only way exchange rates between the dollar and foreign currencies could change was if the surplus nations revalued their currencies upward relative to the dollar. They did not do this frequently enough, so the United States, with its chronically overvalued currency, ran persistent balance of payments deficits in the 1960s.

ADJUSTMENT MECHANISMS UNDER FIXED EXCHANGE RATES

Under the Bretton Woods system of fixed exchange rates, devaluation was viewed as a last resort, to be used only after other methods of adjusting to payments imbalances had failed. What were these other methods?

We have already encountered most of them in our discussion of exchange rate determination in free markets (see pages 823–830). Any factor that increases the demand for, say, U.S. dollars or that reduces the supply will push the value of the dollar upward if it is free to adjust. If, however, the exchange rate is pegged, it is the balance of payments deficit rather than the exchange rate that will adjust. Specifically, the U.S. balance of payments deficit will shrink if either the demand for dollars increases or the supply decreases.

The two panels of Figure 35-7 illustrate this adjustment. In each case, the United States has a payments deficit, since the official exchange rate (3 marks) exceeds the equilibrium rate (2 marks). The deficit starts at *AB* in each diagram. Then either the demand curve moves outward as in Panel (a), or the supply curve moves inward as in Panel (b). With the exchange rate held at 3 marks to the dollar, the balance of payments deficit shrinks—to *CB* in Panel (a) or to *AC* in Panel (b).

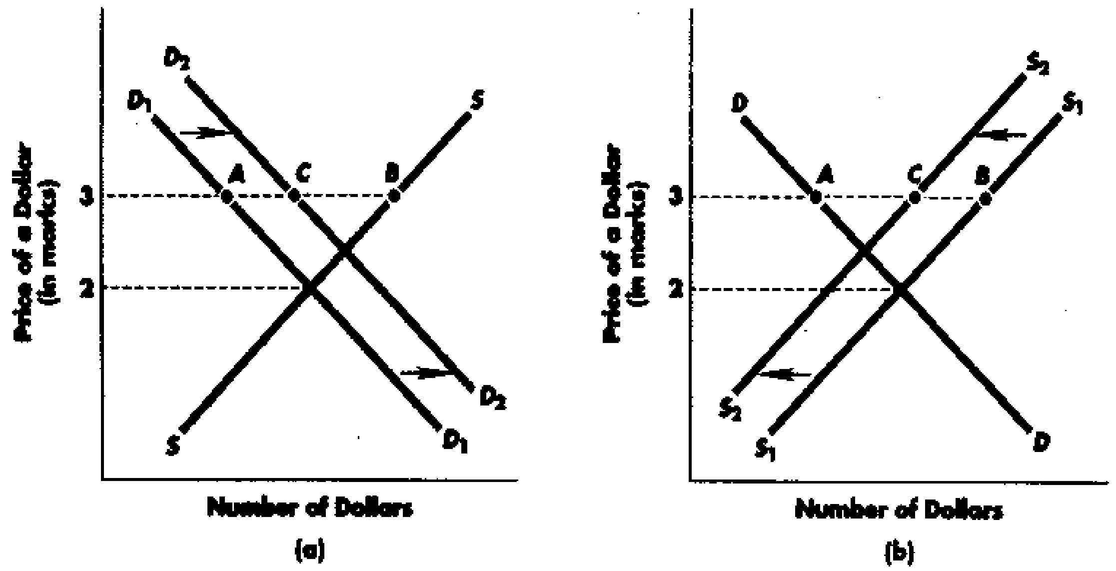
Referring back to our earlier discussions of the factors that underlie the demand and supply curves, then, we see that one way a deficit nation can improve its balance of payments is to *reduce its aggregate demand*, thus discouraging imports and cutting down its demand for foreign currency. Another is to *slow its rate of inflation*, thus encouraging exports and discouraging imports. Finally, it can *raise its interest rates* in order to attract more foreign capital.

In other words, deficit nations were expected to follow restrictive monetary and fiscal policies *voluntarily* just as they would have done *automatically* under the old gold standard. However, just as under the gold standard, this medicine was often unpalatable—as it was to the United States in the 1960s.

A surplus nation could, of course, have taken the opposite measures: pursuing expansive monetary and fiscal policies to increase economic growth and lower interest rates. By increasing the supply of the country's currency and reducing the demand for it, such actions would have reduced the balance of payments surplus. But often surplus countries did not relish the inflation that accompanies expansionary policies; and so, once again, they left the burden of adjustment to the deficit nations. The general point about fixed exchange rates is that:

Under a system of fixed exchange rates, the government of a country loses some control over its domestic economy. There may be times when balance of payments considerations force it to contract its economy in order to cut down its demand for foreign currency, even though domestic needs are calling for expansion. Conversely, there may be times when the domestic economy needs to be reined in, but balance of payments considerations suggest expansion.

FIGURE 35-7 ADJUSTING TO BALANCE OF PAYMENTS DEFICITS



The two panels of this diagram illustrate alternative ways to cut America's balance of payments deficit while maintaining the exchange rate at 3 marks per dollar. Panel (a) might represent a reduction in our inflation rate, which would increase world demand for U.S. export products. Or it could represent a rise in American interest rates, which would attract foreign capital. Panel (b) might represent a reduction in domestic incomes, which would diminish American appetites for foreign goods. In either case, whether demand rises or supply falls, the balance of payments deficit is reduced: from *AB* to *CB* in Panel (a) and from *AB* to *AC* in Panel (b).

While we used the historical example of the dollar-mark exchange rate in Figure 35-7, this discussion also helps us understand what the government of Mexico would have needed to do in late 1994 to defend its exchange rate. But Mexico was in an election year and beset by political instability. So the authorities did not want to risk the recession that more restrictive monetary and fiscal policies probably would have brought on. Instead of taking that bitter medicine, they hoped for the best. But they did not get it.

Back to history. The Bretton Woods system worked fairly well for a number of years, but it finally broke down over its inability to devalue the U.S. dollar. By August 1971, the depletion of America's reserves and the accumulation of foreign debts resulting from America's chronic balance of payments deficits forced President Richard Nixon to end fixed exchange rates, which he did by announcing that the United States would no longer peg the value of the dollar by buying and selling gold. After some futile attempts by the major trading nations to reestablish fixed rates in 1971 and 1972, the Bretton Woods system ended in 1973.

Most observers agree, however, that fixed exchange rates could not have survived the tumultuous events of the next decade in any case. The worldwide inflationary boom of 1972, the supply-side inflations of 1972 to 1974 and 1979 to 1980, and the great worldwide recessions of 1974 to 1976 and the early 1980s led to dramatically different inflation rates in the major countries. For example, between 1975 and 1985, inflation averaged 4 percent per year in Germany, 7 percent in the United States, 11 percent in Great Britain, and 15 percent in Italy. As the purchasing-power parity theory reminds us, large differences in inflation rates call for *major* changes in currency values. The Bretton Woods system was ill-suited to handle such major changes.

WHY TRY TO FIX EXCHANGE RATES?

In view of these and other severe problems with fixed exchange rates, why did the international financial community work so hard to maintain them for so many years? And why do some people today want to return to fixed exchange rates? The answer is that floating exchange rates, determined in free markets by supply and demand, also pose problems.

Chief among these is the possibility that freely floating rates might prove to be highly variable rates, which add an unwanted element of risk to foreign trade. For example, if the exchange rate is 20 cents to the French franc, then a 2,000-franc Parisian dress will cost $400. But should the franc appreciate to 25 cents, that same dress would cost $500. An American department store thinking of buying the dress may need to place its order far in advance and will want to know the cost *in dollars*. It may be worried about the possibility that the value of the franc will rise, so that the dress will cost more than $400. And such worries might inhibit trade.

There are two answers to this concern. First, we might hope that freely floating rates would prove to be fairly stable in practice. Prices of most ordinary goods and services, for example, are determined by supply and demand in free markets and do not fluctuate unduly. Unfortunately, experience since 1973 has dashed this hope. Exchange rates have been extremely volatile—much more volatile than advocates of floating rates anticipated. This volatility is a major reason why some observers want to move back toward fixed exchange rates.

A second possibility is that speculators might relieve business firms of exchange rate risks—for a fee, of course. Consider the department store example. If French francs cost 20 cents today, the department store manager can assure herself of paying exactly $400 for the dress several months from now by arranging for a speculator to deliver francs to her at 20 cents per franc on the day she needs them. If the franc appreciates in the interim, it is the speculator, not the department store, that will take the financial beating. And, of course, if the franc depreciates, the speculator will pocket the profits. Thus, speculators play an important role in a system of floating exchange rates.



"Then it's agreed. Until the dollar firms up, we let the clamshell float."

SOURCE: Drawing by Ed Fisher © 1971, The New Yorker Magazine, Inc.

The widespread fears that speculative activity in free markets will lead to wild gyrations in prices, while occasionally valid, are often unfounded. The reason is simple. To make profits, international currency speculators must buy a currency when its value is low (thus helping to support the currency by pushing up its demand curve) and sell it when its value is high (thus holding down the price by adding to the supply curve).

This means that, to be successful, speculators must come into the market as *buyers* when demand is weak (or when supply is strong), and come in as *sellers* when demand is strong (or supply is scant). In doing so, they will help limit price fluctuations. Looked at the other way around, speculators can destabilize prices only if they are systematically willing to lose money.[5]

Notice the stark contrast to the system of fixed exchange rates in which speculation often leads to wild "runs" on currencies that are on the verge of devaluation. Speculative activity, which may well be destabilizing under fixed rates, is more likely to be stabilizing under floating rates.

[5]See Review Question 11 at the end of the chapter.

We do not mean to imply that speculation makes floating rates trouble-free. At the very least, speculators will demand a fee for their services—a fee that adds to the costs of trading across national borders. In addition, not all exchange rate risks can be eliminated through speculation. For example, few contracts on foreign currencies nowadays last more than, say, a year or two. Thus, a business cannot easily protect itself from exchange rate changes over periods of many years. Finally, speculative markets can and do get carried away from time to time, moving currency rates in ways that are hard to understand and that governments find frustrating.

Yet, despite all these problems, international trade has flourished under floating exchange rates. Apparently, exchange rate risk is not as burdensome as some people feared.

THE CURRENT MIXED SYSTEM

Our current international financial system—where some currencies are still pegged in the old Bretton Woods manner, many are floating freely, and others are floating subject to government interferences—has evolved gradually since the United States severed the dollar's link to gold. Though it continues to change and adapt, at least three features have been evident.

The first is the decline in the notion that exchange rates should be fixed for long periods of time. The demand by many countries that the world quickly return to fixed exchange rates had largely subsided by the mid-1970s. Even where rates are still pegged, devaluations and revaluations are now much more frequent—and smaller—than they were in the Bretton Woods era. Most currency rates change slightly on a day-to-day basis, and market forces generally determine the basic trends, up or down. Even advocates of greater fixity in exchange rates generally propose that governments keep rates within certain *ranges*, rather than literally fix them.

Second, however, central banks do not hesitate to intervene to moderate exchange movements whenever they feel that such actions are appropriate. Typically, these interventions are aimed at ironing out what are deemed to be transitory fluctuations. But there are times in which central banks oppose basic trends in exchange rates. For example, the Federal Reserve and other central banks sold dollars aggressively in 1985 to push the dollar down, and they bought dollars in 1994 and 1995 to arrest the rise of the mark and, especially, the yen. While the world has relatively few truly fixed exchange rates, most of the major currencies are floating less than freely. The terms *dirty float* or *managed float* have been coined to describe this mongrel system.

The third unmistakable feature of the present international monetary system is the virtual elimination of any role for gold. The trend away from gold actually began before 1971, and by now gold plays essentially no role in the world's financial system. Instead, there is a *free market* in gold in which dentists, jewelers, industrial users, speculators, and ordinary citizens who think of gold as a good store of value can buy or sell as they wish. The price of gold is determined each day by the law of supply and demand.

THE VOLATILE DOLLAR

We mentioned earlier that floating exchange rates have not been stable exchange rates. No currency illustrates this better than the U.S. dollar. (See Figure 35-8.)

FIGURE 35-8 THE UPS AND DOWNS OF THE DOLLAR



This graph charts the behavior of the international value of the dollar relative to a basket of ten major foreign currencies since 1974. (The index is based on March 1973 = 100.) The net change in the value of the dollar over the entire period is small, but the ups and downs have been pronounced. The stunning climb of the dollar from its 1980 low to its 1985 high stands out on the graph, as does the pronounced decline from 1985 to 1988.

SOURCE: Federal Reserve System.

In 1977 and 1978, the international value of the dollar plummeted until a concerted effort by central banks to buy dollars stopped the fall. The dollar then stabilized for almost 2 years before rising like a rocket for a period of almost 5 years. As Table 35-1 (page 823) showed, in July 1980 a U.S. dollar bought less than 2 German marks, about 4 French francs, and about 830 Italian lira. By the time it peaked in February 1985, the mighty dollar could buy more than 3 German marks, about 10 French francs, and over 2,000 Italian lira. Such major currency changes had dramatic effects on world trade.

The rising dollar was a blessing to Americans who traveled abroad or who bought foreign goods because foreign prices, when translated to dollars by the exchange rate, looked cheap to Americans.[6] But the arithmetic worked just the other way for U.S. firms seeking to sell their goods abroad; foreign buyers found everything American very expensive.[7] It was no surprise, therefore, that as the dollar climbed our exports fell, our imports rose, and many of our leading manufacturing industries were decimated by foreign competition. An expensive currency, Americans came to learn, is a mixed blessing.

---

[6]EXAMPLE: How much does a 600-franc hotel room in Paris cost in dollars when the franc is worth 20 cents? 16 cents? 10 cents?

[7]EXAMPLE: How much does a $55 American camera cost a German consumer when the mark is worth 55 cents? 44 cents? 33.33 cents?

From early 1985 until early 1988, the value of the dollar fell even faster than it had risen. The cheaper dollar curbed American appetites for imports and alleviated the plight of our export industries, many of which boomed. However, rising prices for imported goods and foreign vacations were a source of consternation to many American consumers.

Since 1988, the overall value of the dollar has not changed much—on average. But there have been quite a few appreciations and depreciations sharp enough to gain the attention of the world's financial markets and central banks. The international currency markets are anything but boring.

THE EUROPEAN EXCHANGE RATE MECHANISM

As noted earlier, floating exchange rates are no magical cure-all. One particular problem confronts the members of the European Union (EU). These countries have pledged to create a unified market like the United States. As part of that pledge, they need eventually to establish a single currency for all member countries. That, of course, precludes floating rates.

The *European Exchange Rate Mechanism (ERM)* was designed as an important step along the road to fixed exchange rates within Europe, and thence to a common currency. Under the ERM, the values of the major European currencies are supposed to remain fixed within relatively narrow bands, while EU currencies as a group rise or fall *relative to the rest of the world*. The German mark has clearly become the dominant currency in this group—so much so that many observers now see Europe as a "mark area."

The ERM has had a tumultuous history thus far in the 1990s. Problems first arose in 1990 when the reunification of Germany required a variety of major economic adjustments—including, perhaps, a change in the value of the mark. Germany found itself in somewhat the same position as the United States under Bretton Woods: Since the mark could not easily be realigned within the ERM, the burden of adjustment fell on other countries.

In 1992, the system began to break down, causing great currency turmoil throughout Europe. The United Kingdom and Italy left the ERM and several other countries realigned their currencies vis-à-vis the mark. But, despite tremendous pressure, the French franc maintained its fixed parity with the mark and the ERM weathered the storm—albeit smaller and more fragile than before.

The calm lasted less than a year. In 1993, speculators became convinced that several European currencies would have to be devalued relative to the mark; and they began selling these currencies to buy marks. Europe's central banks resisted for a while, but eventually succumbed to market forces. Currency bands were widened enough to give the weaker currencies room to float downward, and with that added flexibility the storm passed with the ERM intact. Since then, the system has held together, albeit without the pound and the lira, despite periodic attacks by speculators.

But many observers have come to question how and when the next step—to a common currency—will take place. The EU countries are committed to this goal under the Treaty of Maastricht, and 1999 is the target date. But as of 1996 there were many clouds on the horizon. The United Kingdom and Italy were still out of the system. Almost none of the countries met the fiscal requirements of the Maastricht treaty. And public opinion in many countries opposed the move to a common currency. On balance, prospects for a single currency by 1999 looked shaky.

SUMMARY

1. **Exchange rates** state the value of one currency in terms of another and thus influence the patterns of world trade in important ways.
2. If governments do not interfere by buying or selling their currencies, exchange rates will be determined in free markets by the usual laws of supply and demand. Such a system is called **floating exchange rates.**
3. Demand for a nation's currency is derived from foreigners' desires to purchase that country's goods and services or to invest in its assets. Any change that increases the demand for a nation's currency will cause its exchange rate to **appreciate** under floating rates.
4. Supply of a nation's currency is derived from the desire of that country's citizens to purchase foreign goods and services or to invest in foreign assets. Any change that increases the supply of a nation's currency will cause its exchange rate to **depreciate** under floating rates.
5. In the long run, purchasing-power parity plays a major role in exchange rate movements. The *purchasing-power parity theory* states that relative price levels in any two countries determine the exchange rate between their currencies. Therefore, countries with relatively low inflation rates normally will have appreciating currencies.
6. Over shorter periods, purchasing-power parity has little influence over exchange rate movements. The pace of economic activity and the level of interest rates exert greater influence. In particular, short-term capital movements are typically the dominant factor.
7. Exchange rates can be fixed at nonequilibrium levels by governments that are willing and able to mop up any excess of quantity supplied over quantity demanded, or provide any excess of quantity demanded over quantity supplied. In the first case, the country is suffering from a **balance of payments deficit** because of its overvalued currency. In the second, an undervalued currency has given it a **balance of payments surplus.**
8. In the early part of this century, the world was on a particular system of **fixed exchange rates** called the *gold standard*, in which the value of every nation's currency was fixed in terms of gold. But this created problems because nations could not control their own money supplies and because the world could not control its total supply of gold.
9. After World War II, the gold standard was replaced by the *Bretton Woods system* under which rates were again fixed—or rather fixed until further notice. This system made the U.S. dollar the basis of international currency values.
10. The Bretton Woods system served the world well and helped restore world trade, but it ran into trouble when the dollar became chronically overvalued since the system provided no way to remedy this situation.
11. Since 1971, the world has moved toward a system of relatively free exchange rates, though there are plenty of exceptions. We now have a thoroughly mixed system of *"dirty"* or *"managed" floating* which continues to evolve and adapt.
12. Floating rates are not without their problems. For example, importers and exporters justifiably worry about fluctuations in exchange rates. Though these problems seem manageable, some people would like to see a return to fixed exchange rates.
13. Under floating exchange rates, investors who speculate on international currency values provide a valuable service by assuming the risks of those who do not wish to speculate. Normally, speculators stabilize rather than destabilize exchange rates, because that is how they make profits.
14. The U.S. dollar rose dramatically in value from 1980 to 1985, making our imports cheaper and our exports more expensive. Then, from 1985 to 1988, the dollar tumbled, which had precisely the reverse effects.
15. The European Union seeks eventually to establish a single currency for all member nations. As a step in that direction, the European Exchange Rate Mechanism was set up to fix exchange rates among member countries.

KEY TERMS

International monetary system
Exchange rate
Appreciation
Depreciation
Devaluation
Revaluation
Supply of and demand for foreign exchange
Floating exchange rates
Purchasing-power parity theory
Fixed exchange rates

Balance of payments deficit and surplus
Current account
Capital account
Gold standard
Bretton Woods system
International Monetary Fund (IMF)
"Dirty" or "managed" float
European Exchange Rate Mechanism (ERM)

QUESTIONS FOR REVIEW

1. What items do you own or routinely consume that are produced abroad? From what countries do these come? Suppose Americans decided to buy more of these things? How would that affect the exchange rates between the dollar and these currencies?
2. If the dollar depreciates relative to the Japanese yen, will the Sony television you have wanted become more or less expensive? What effect do you imagine this will have on American demands for Sonys? Does the American demand curve for yen, therefore, slope upward or downward? Explain.
3. During the first half of the 1980s, inflation in (West) Germany was consistently lower than that in the United States. What, then, does the purchasing-power parity theory predict should have happened to the exchange rate between the mark and the dollar between 1980 and 1985? (Look at Table 35-1 to see what actually happened.)
4. Use supply and demand diagrams to analyze the effect on the exchange rate between the dollar and the yen if:
   a. Japan opens its domestic markets to more foreign competition.
   b. Investors come to fear that values on the Tokyo stock market will decline.
   c. The Federal Reserve lowers interest rates in America.
   d. The U.S. government, to help settle the problems of the Middle East, gives huge amounts of foreign aid to Israel and her Arab neighbors.
   e. Both Japan and the United States recover from recessions, but the Japanese recovery is slower.
   f. Prospects for inflation in the United States worsen.
5. How are the problems of a country faced with a balance of payments deficit similar to those posed by a government regulation that holds the price of milk above the equilibrium level? (*Hint:* Think of each in terms of a supply-demand diagram.)
6. For each of the transactions listed below, indicate how it would affect the U.S. balance of payments if exchange rates were fixed.
   a. You spent the summer traveling in Europe.
   b. Your uncle in Canada sent you $20 as a birthday present.
   c. You bought a new Honda.
   d. You sold some stock you own on the Tokyo Stock Exchange.
7. Suppose each of the transactions listed in Review Question 6 was done by many Americans. Indicate how each would affect the international value of the dollar if exchange rates were floating.
8. Under the old gold standard, what do you think happened to world prices when there was a huge gold strike in California in 1849? What do you think happened when the world went without any important new gold strikes for 20 years or so?
9. Explain why the members of the Bretton Woods conference in 1944 wanted to establish a system of fixed exchange rates. What was the flaw that led to the ultimate breakdown of the system in 1971?
10. Suppose you want to reserve a hotel room in London for the coming summer but are worried that the value of the pound may rise between now and then, making the rooms too expensive for your budget. Explain how a speculator could relieve you of this worry. (Don't actually try it. Speculators deal only in very large sums!)
11. We learned on page 838 that successful speculators buy a currency when demand is weak and sell it when demand is strong. Use supply and demand diagrams for two different periods (one with weak demand, the other with strong demand) to show why this will limit price fluctuations.
12. In early 1995, market forces were pushing up the international value of the Japanese yen. Why do you think the government of Japan was unhappy about this currency appreciation? What could the Bank of Japan (Japan's central bank) have done to try to prevent it? How could the Federal Reserve have helped? Why might the central banks have failed in this attempt?

CHAPTER 36

MACROECONOMICS IN A WORLD ECONOMY

No man is an island,
entire of itself.
John Donne

America is not an isolated economy immune from foreign influences. Today, more than ever before, the nations of the world are locked together in an uneasy economic union. Fluctuations in foreign GDP growth, foreign inflation, and foreign interest rates profoundly affect the U.S. economy. Economic events that originate in our country reverberate around the globe. Without a deeper understanding of these international linkages, we cannot hope to understand many of the most important economic developments of our time.

What we have learned so far about the macroeconomics of international trade in goods and services was correct, but limited. In particular, it ignored such crucial influences as exchange rates and international financial movements. Changes in exchange rates alter the prices of one country's goods in terms of the currency of another. In the previous chapter, we learned how major macroeconomic variables such as GDP, prices, and interest rates affect exchange rates. In this chapter, we complete the circle by studying how changes in the exchange rate affect the domestic economy. Then we bring international capital flows into the picture and learn how monetary and fiscal policy work in an **open economy.**

An **OPEN ECONOMY** is one that trades with other nations in goods and services, and perhaps also in financial assets.

ISSUE: DEFICIT REDUCTION AND THE DOLLAR

When the movement to balance the federal budget caught political fire in 1995, a lively side-debate arose over how lower deficits would affect the value of the dollar. Many politicians and business people argued that deficit reduction would strengthen the dollar. In their view, decisive action to cut government spending would bolster confidence in the United States, thereby increasing the international demand for dollars and making the dollar appreciate.

Most economists, however, took the opposite position, arguing that lower deficits should lead to a lower dollar. As evidence, they pointed to the events of the early 1980s, when soaring U.S. budget deficits led to a stunning appreciation of the dollar. Shouldn't the opposite happen, they asked, if the budget deficit is cut? In this chapter, we will learn the reasoning behind the economists' position.

INTERNATIONAL TRADE AND AGGREGATE DEMAND: A QUICK REVIEW

We know from earlier chapters (especially Chapters 25 and 26) that a country's net exports, $(X - IM)$, are one component of its aggregate demand, $C + I + G + (X - IM)$. For this reason, an autonomous increase in exports or decrease in imports has a multiplier effect on the economy, just like an increase in consumption, investment, or government purchases.[1] Figure 36-1 depicts this conclusion on an aggregate demand and supply diagram. A rise in net exports shifts the aggregate demand curve outward to the right, thereby raising both GDP and the price level.

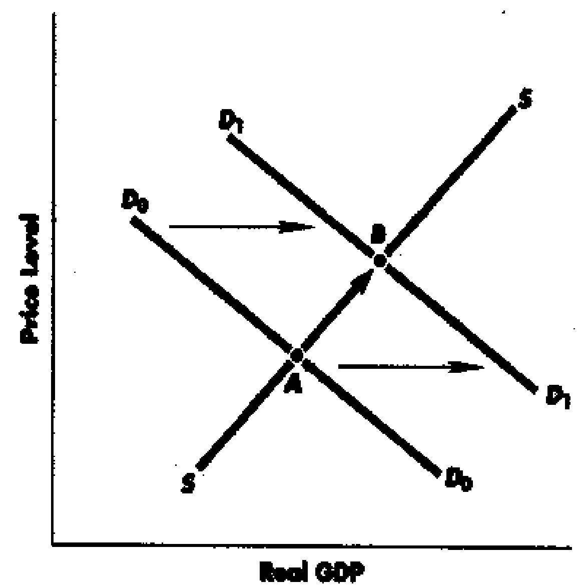
But what can make net exports rise? One factor we mentioned in Chapter 25 was a rise in foreign incomes. If foreigners become richer, they are likely to spend more on a wide variety of products, some of which will be American exports. Thus, Figure 36-1 illustrates the effect on the U.S. economy of a boom in foreign countries. Similarly, a recession abroad would reduce U.S. exports and shift the U.S. aggregate demand curve inward. Thus, as we learned in Chapter 26:

Booms or recessions in one country tend to be transmitted to other countries through international trade in goods and services.

One other important determinant of net exports was mentioned in Chapter 25, but not discussed in depth: the relative prices of foreign and domestic goods. The idea is a simple application of the law of demand: If the prices of the goods of Country X rise, then people everywhere will tend to buy fewer of them—and

FIGURE 36-1 THE EFFECTS OF HIGHER NET EXPORTS

If real exports rise or real imports fall, the economy's aggregate demand curve shifts outward, from D_0D_0 to D_1D_1. Real GDP and the price level both rise as the equilibrium moves from point A to point B.



[1]An appendix to Chapter 26 showed that international trade lowers the numerical value of the multiplier. Nonetheless, autonomous changes in C, I, G, or $(X - IM)$ all have the same multiplier.

more of the goods of Country Y. As we shall see now, this simple idea holds the key to understanding how exchange rates affect international trade.

RELATIVE PRICES, EXPORTS, AND IMPORTS

First assume—just for this short section—that exchange rates are *fixed.* What happens if the prices of American goods fall while, say, Japanese prices are constant? With U.S. products now less expensive *relative to Japanese products,* both Japanese and American consumers will probably buy more American goods and fewer Japanese goods. So America's exports will rise and imports will fall, adding to aggregate demand in this country—as shown in Figure 36-1. Conversely, a rise in American prices (relative to Japanese prices) will *decrease* our net exports and aggregate demand. Thus:

For given foreign prices, a fall in the prices of a country's exports will lead to an increase in that country's net exports, and hence to a rise in its real GDP. Analogously, a rise in the prices of a country's exports will decrease that country's net exports and GDP.

Precisely the same logic applies to changes in Japanese prices. If Japanese prices rise, Americans will export more and import less. So $X - IM$ will rise, boosting GDP in the United States. Figure 36-1 applies to this case without change. By similar reasoning, falling Japanese prices decrease U.S. net exports and depress our economy. Thus:

Price increases abroad raise a country's net exports and hence its GDP. Price decreases abroad have the opposite effects.

THE EFFECTS OF CHANGES IN EXCHANGE RATES

From here it is a simple matter to figure out how changes in *exchange rates* affect a country's net exports, for currency appreciations or depreciations change international relative prices.

Recall that the basic role of an exchange rate is to convert one country's prices into the currency of another. Table 36-1 uses two examples of U.S.-Japanese trade to remind us of this role. Suppose the dollar depreciates from 120 yen to 100 yen.

TABLE 36-1 EXCHANGE RATES AND HOME-CURRENCY PRICES

	¥60,000 Japanese TV Set		$2,000 U.S. Home Computer	
Exchange Rate	**Price in Japan**	**Price in United States**	**Price in United States**	**Price in Japan**
$1 = 120 yen	¥60,000	$500	$2,000	¥240,000
$1 = 100 yen	¥60,000	$600	$2,000	¥200,000

Then, from the viewpoint of American consumers, a television set that costs ¥60,000 in Japan goes up in price from $500 (that is, 60,000/120) to $600. To Americans, it is just as if Japanese manufacturers had raised TV prices by 20 percent. Naturally, Americans react by purchasing fewer Japanese products. So American imports go down.

Now consider the implications for Japanese consumers interested in buying American personal computers that cost $2,000. When the dollar falls from 120 yen to 100 yen, they see the price of these computers falling from ¥240,000 (that is, 2,000 × 120) to ¥200,000. To them, it is just as if American producers had offered a 16.7 percent markdown. Under such circumstances, we expect U.S. sales to the Japanese to rise. So U.S. exports should increase. Putting these two findings together, we conclude that:

A currency depreciation should raise net exports and therefore increase aggregate demand. Conversely, a currency appreciation should reduce net exports and therefore decrease aggregate demand.

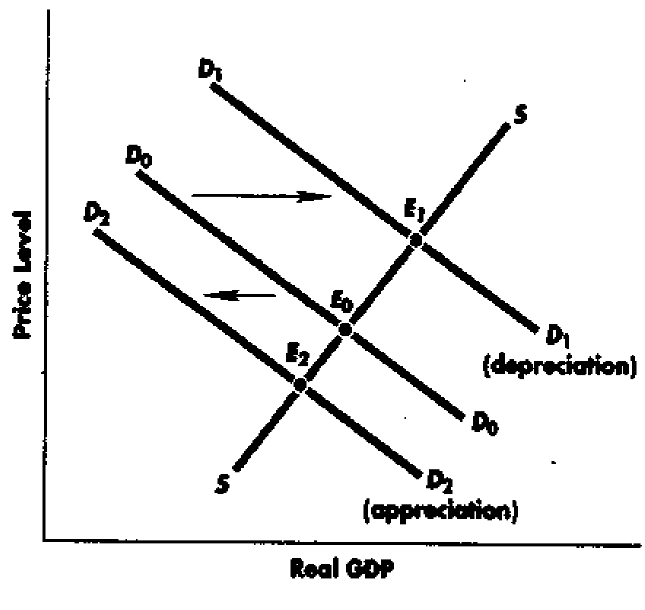
The aggregate supply and demand diagram in Figure 36-2 illustrates this conclusion. If the currency depreciates, net exports rise and the aggregate demand curve shifts outward from D_0D_0 to D_1D_1. Both prices and output rise as the economy's equilibrium moves from E_0 to E_1. If the currency appreciates, everything operates in reverse: Net exports fall, the aggregate demand curve shifts inward to D_2D_2, and both prices and output decline.

This simple analysis already helps us understand why the U.S. trade deficit grew so large in the 1980s. We saw in Figure 35-8 (page 840) that the international value of the dollar soared in the first half of the 1980s. According to the reasoning we have just completed, such a stunning appreciation of the dollar should have raised U.S. imports and damaged U.S. exports. That is precisely what happened. In constant dollars, American imports soared by 60 percent

FIGURE 36-2

THE EFFECTS OF EXCHANGE RATE CHANGES ON AGGREGATE DEMAND

A depreciation of the exchange rate raises net exports and hence shifts the aggregate demand curve outward to the right, from D_0D_0 to D_1D_1 in the diagram. An appreciation of the currency shifts the aggregate demand curve inward to the left, to D_2D_2. Thus, depreciations are expansionary and appreciations are contractionary.



between 1981 and 1986, while American exports rose a scant 8 percent. The result is that a $6 billion net export *surplus* in 1981 turned into a $164 billion *deficit* by 1986.

AGGREGATE SUPPLY IN AN OPEN ECONOMY

So we have concluded that a currency depreciation increases aggregate demand while a currency appreciation decreases it. To complete our model of macroeconomics in an open economy, we must now turn to the implications of international trade for *aggregate supply.*

Part of the story is familiar. As we know from previous chapters, the United States, like all economies, purchases some of its productive inputs from abroad. Oil is only the most prominent example. We also rely on foreign suppliers for metals like titanium, raw agricultural products like coffee beans, and thousands of other items used by American industry. When the dollar depreciates, these imported inputs become more costly in terms of U.S. dollars—just as if foreign prices had risen.

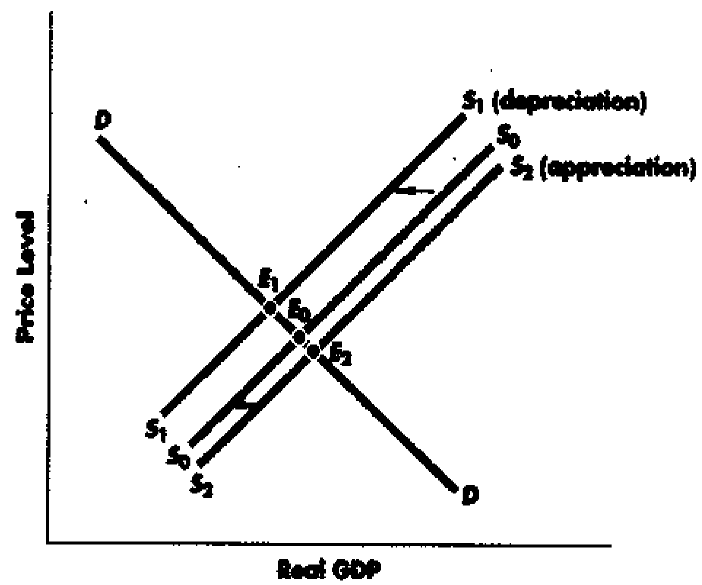
The consequence is clear: With imported inputs more expensive, American firms will be forced to charge higher prices at any given level of output. Graphically, this means that *the aggregate supply curve will shift inward* (to the left).

When the dollar depreciates, the prices of imported inputs rise. The U.S. aggregate supply curve therefore shifts inward, pushing up the prices of American-made goods and services. By exactly analogous reasoning, an appreciation of the dollar makes imported inputs cheaper and shifts the U.S. aggregate supply curve *outward,* thus pushing American prices down. (See Figure 36-3.)

Beyond this, a depreciating dollar has further inflationary effects that do not show up on the aggregate demand and supply diagram. Most obviously, prices

FIGURE 36-3

THE EFFECTS OF EXCHANGE RATE CHANGES ON AGGREGATE SUPPLY



A depreciation of the currency pushes the aggregate supply curve inward, from S_0S_0 to S_1S_1 in the diagram, and is therefore inflationary. A currency appreciation has a deflationary effect because it pushes the aggregate supply curve outward, from S_0S_0 to S_2S_2.

of imported goods are included in American price indexes like the Consumer Price Index (CPI). So when dollar prices of Japanese cars, French wine, and Swiss watches increase, the CPI goes up *even if no American prices rise.* For this and other reasons, the inflationary impact of a dollar depreciation is even greater than that indicated by Figure 36-3.[2]

THE MACROECONOMIC EFFECTS OF EXCHANGE RATES

Let us now put aggregate demand and aggregate supply together and study the macroeconomic effects of changes in exchange rates.

First suppose that the international value of the dollar falls. Referring back to Figures 36-2 and 36-3, we see that this will shift the aggregate demand curve *outward* and the aggregate supply curve *inward*. The result, as Figure 36-4 shows, is that the U.S. price level certainly rises. Whether real GDP rises or falls depends on whether the supply or demand shift is the dominant influence. The evidence strongly suggests that aggregate *demand* shifts are usually more important, so we expect GDP to rise. Hence:

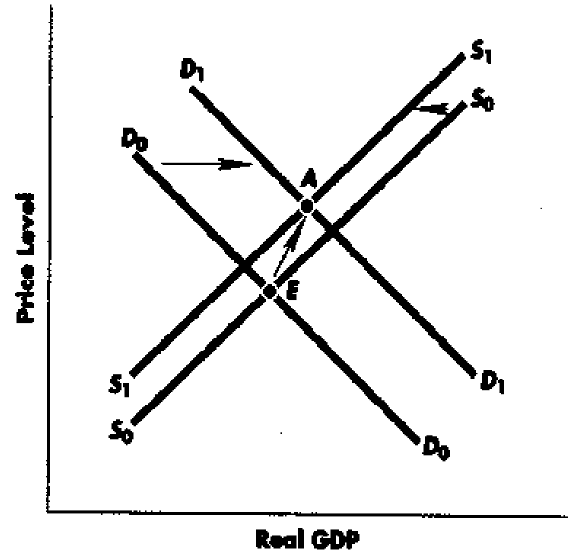
A currency depreciation is inflationary and probably also expansionary.

What is the intuitive explanation for this result? When the dollar falls, foreign goods become more expensive to Americans. That is directly inflationary. At the same time, aggregate demand in the United States is stimulated by rising net exports. As long as the expansion of demand outweighs the adverse shift of the aggregate supply curve brought on by the depreciation, real GDP should rise.

Now let's reverse things. Suppose the dollar *appreciates*. In this case, net exports *fall* so the aggregate demand curve shifts *inward*. At the same time,

FIGURE 36-4 THE EFFECTS OF A CURRENCY DEPRECIATION

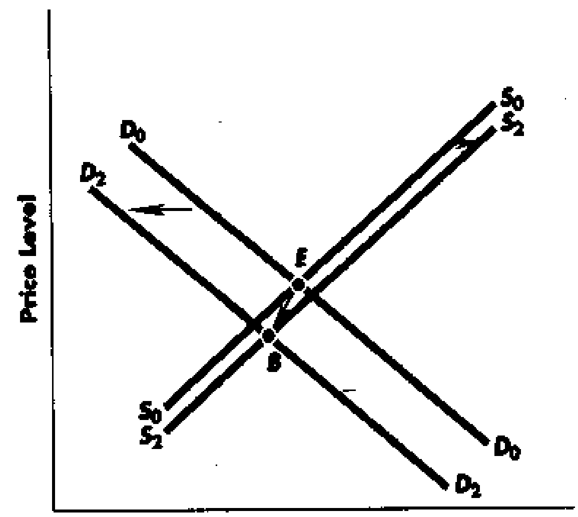
If the currency depreciates, aggregate demand increases because net exports are stimulated, and aggregate supply decreases because imported inputs become more expensive. Prices rise as equilibrium moves from point *E* to point *A*. If the demand shift is the more important influence, output increases, too.



[2]The diagram should be interpreted as showing the effects of currency depreciations and appreciations on the prices of *domestically produced* goods.

FIGURE 36-5 THE EFFECTS OF A CURRENCY APPRECIATION

If the currency appreciates, aggregate demand declines because net exports fall, and aggregate supply increases because imported inputs become cheaper. Prices fall as equilibrium moves from point *E* to point *B*. If the demand shift is the more important influence, output also falls.



imported inputs become cheaper, so the aggregate supply curve shifts *outward*. (See Figure 36-5.) Once again, we can be sure of the movement of the price level: It falls. Output also falls if the demand shift is more important than the supply shift, as is likely. Thus:

A currency appreciation is certainly disinflationary and is probably contractionary.

In the rest of the chapter, we shall assume that the effect of the exchange rate on aggregate *demand* dominates its effect on aggregate *supply*, because that is what the evidence suggests.

INTEREST RATES AND INTERNATIONAL CAPITAL FLOWS

There is one important piece left in our international economic puzzle. We have analyzed international trade in goods and services rather fully, but we have ignored international movements of *capital*.

For some nations, this omission is of little consequence because they are rarely involved in international capital flows. But things are quite different for the United States because the dollar is the world's major international currency. The vast majority of international financial flows involve either buying or selling assets whose values are stated in U.S. dollars. Fortunately, given what we have just learned about the effects of exchange rates, it is easy to add international capital flows to our analysis.

Recall from the previous chapter that interest rate differentials and capital flows are typically the most important determinants of exchange rate movements in the short run. Specifically, suppose interest rates in the United States rise while foreign interest rates remain unchanged. We learned in the last chapter that this will attract capital to the United States and cause the dollar to appreciate. This chapter has just taught us that an appreciating dollar will, in turn, reduce net

exports, prices, and output in the United States—as indicated in Figure 36-5. Thus:

A rise in interest rates tends to contract the economy by appreciating the currency and reducing net exports.

Notice that this conclusion has a familiar ring. While studying monetary policy in Chapter 30, we observed that higher interest rates deter investment spending and hence reduce the I component of $C + I + G + (X - IM)$. Now, in studying an open economy with international capital flows, we see that higher interest rates also reduce the $X - IM$ component. Thus, *international capital flows strengthen the negative effects of interest rates on aggregate demand.*

If interest rates in the United States fall, or if those abroad rise, everything we have just said is turned in the opposite direction. The conclusion is:[3]

A decline in interest rates tends to expand the economy by depreciating the currency and raising net exports.

FISCAL POLICY IN AN OPEN ECONOMY

Now we are ready to use our model to analyze how fiscal and monetary policy work when the exchange rate is floating and capital is internationally mobile. Doing so will teach us how international economic relations modify the effects of stabilization policies that we learned in earlier chapters. Fortunately, no new theoretical apparatus is necessary; we only need remember what we have learned in the chapter up to this point. Specifically:

- A rise in the domestic interest rate leads to capital inflows and makes the exchange rate appreciate. A fall in the domestic interest rate leads to capital outflows and makes the exchange rate depreciate.
- A currency appreciation reduces aggregate demand and raises aggregate supply (see Figure 36-5). A currency depreciation raises aggregate demand and reduces aggregate supply (see Figure 36-4).

A **CLOSED ECONOMY** is one that does not trade with other nations in either goods or assets.

With this in mind, suppose the government cuts taxes or raises spending. Aggregate demand increases, which pushes up both real GDP and the price level in the usual manner. This is shown as the shift from D_0D_0 to the blue line D_1D_1 in Figure 36-6. In a **closed economy,** that is the end of the story. But in an *open economy* with international capital flows, we must add in the macroeconomic effects that work through the exchange rate. We do this by answering two questions.

First, what will happen to the exchange rate? We know from earlier chapters that a fiscal expansion pushes up interest rates—a fact that is sure to be noticed in international financial markets. At higher interest rates, American securities become more attractive to foreign investors, who go to the foreign exchange markets to buy dollars for use in purchasing American securities. This buying pressure drives up the value of the dollar. Thus:

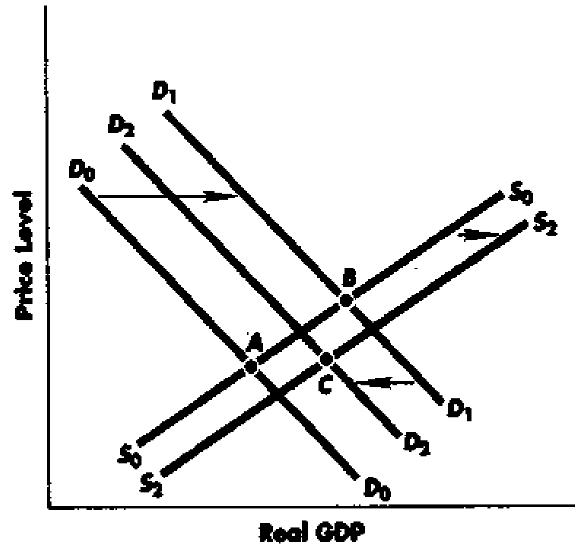
A fiscal expansion normally makes the exchange rate appreciate.

Second, what are the effects of a higher dollar? As we know, when the dollar rises in value, American goods become more expensive abroad and foreign

[3]EXERCISE: Provide the reasoning behind this conclusion.

FIGURE 36-6 A FISCAL EXPANSION IN AN OPEN ECONOMY

A fiscal expansion pushes the aggregate demand curve outward, from D_0D_0 to D_1D_1 in the diagram. But it also raises interest rates, which attracts international capital and appreciates the currency. The currency appreciation, in turn, reduces aggregate demand and raises aggregate supply—as shown by the orange curves S_2S_2 and D_2D_2. The result is that equilibrium occurs at point C rather than at point B. Output and prices both rise less than they would in a closed economy.



goods become cheaper here. So exports fall and imports rise, driving down the $X - IM$ component of aggregate demand. The fiscal expansion thus winds up increasing both America's *capital account surplus* (by attracting foreign capital) and its *current account deficit* (by reducing net exports). In fact, the two must rise by equal amounts because, under floating exchange rates, it is always true that:[4]

$$\text{Current account surplus} + \text{Capital account surplus} = 0$$

Since a fiscal expansion leads in this way to a trade deficit, many economists believe that the U.S. trade deficit of the 1980s was a side effect of the large tax cuts made early in the decade. We will come back to that issue shortly.

But first note that the induced rise in the dollar will shift the aggregate supply curve *outward* and the aggregate demand curve *inward*, as we saw in Figure 36-5. Figure 36-6 adds these two shifts (in orange) to the effect of the original fiscal expansion (in blue). The final equilibrium in an open economy is point C, whereas in a closed economy it would be point B. By comparing points B and C, we can see how international linkages change the picture of fiscal policy that we painted earlier in the book.

There are two main differences. First, a higher exchange rate offsets part of the inflationary effect of a fiscal expansion by making imports cheaper. Second, a higher exchange rate reduces the expansionary effect on real GDP by reducing $X - IM$. Here we have a new kind of "crowding out," different from the one we studied in Chapter 32. There we learned that an increase in G will crowd out some private investment spending by raising interest rates. Here a rise in G, by raising both interest rates and the exchange rate, crowds out *net exports*. But the effect is the same: The fiscal multiplier is reduced. Thus, we conclude that:

International capital flows reduce the power of fiscal policy.

[4] If you need review, turn back to Chapter 35, pages 833–834.

TABLE 36-2

PERCENTAGE SHARES OF REAL GDP
IN THE UNITED STATES: 1981 AND 1986

Year	C	I	G	X − IM
1981	64.5%	14.6%	20.1%	0.1%
1986	67.6%	14.8%	20.7%	−3.0%
Change	+3.1	+0.2	+0.6	−3.1

Note: Totals do not add up due to rounding and deflation.

Table 36-2 suggests that this new international variety of crowding out was much more important than the traditional type of crowding out in the early 1980s. Between 1981 and 1986, the share of investment in GDP actually *increased* slightly (from 14.6 percent to 14.8 percent) despite the rise in the shares of both consumer spending and government purchases. Only the share of net exports, $X - IM$, fell—from 0.1 percent to −3.0 percent.

This was an important lesson that American economists learned in the 1980s. In 1980, many economists worried that large government budget deficits would crowd out private investment. By the end of the decade, most were more concerned that deficits were crowding out net exports. Similarly, nowadays most economists believe that the benefits of a large deficit-reduction program, such as the one Congress and the president debated in 1995 and 1996, should be split between higher investment and a smaller trade deficit.

MONETARY POLICY IN AN OPEN ECONOMY

Now let us consider how monetary policy works in an open economy with floating exchange rates and international capital mobility. To remain consistent with the history of the 1980s, we consider a tightening, rather than a loosening, of monetary policy.

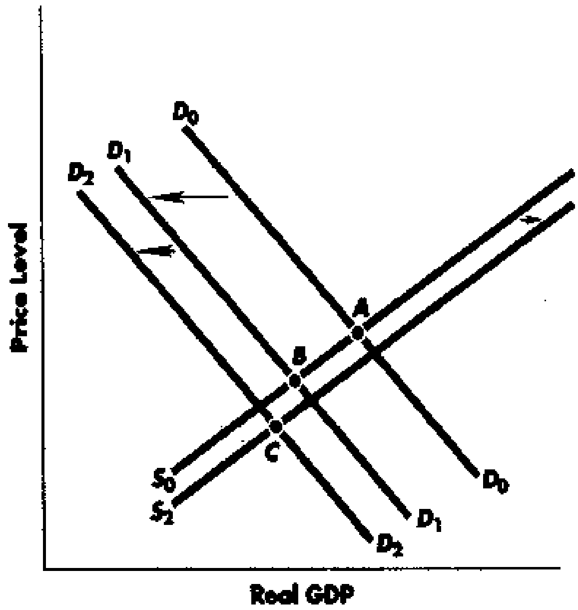
As we know from earlier chapters, contractionary monetary policy reduces aggregate demand, which lowers both real GDP and prices. This is shown in Figure 36-7 by the shift from D_0D_0 to the blue line D_1D_1, and it looks like the exact opposite of a fiscal expansion. Without international capital flows, that would be the end of the story.

But, in the presence of internationally mobile capital, we must think through the consequences for interest rates and exchange rates. We know from previous chapters that a monetary contraction raises interest rates, just like a fiscal expansion. Hence, tighter money attracts foreign capital into the United States in search of higher rates of return. The exchange rate therefore rises. The appreciating dollar encourages imports and discourages exports; so $X - IM$ falls. America therefore winds up with capital flowing in and an increase in its trade deficit. This time, as you will notice from Figure 36-7:

International capital flows increase the power of monetary policy.

Why do capital flows *strengthen* monetary policy but *weaken* fiscal policy? The answer lies in their effects on interest rates. The main international repercussion of either a fiscal *expansion* or a monetary *contraction* is to raise interest rates and

FIGURE 36-7 A MONETARY CONTRACTION IN AN OPEN ECONOMY



A monetary contraction pulls the aggregate demand curve inward, from D_0D_0 to D_1D_1, just like a fiscal contraction. But it raises, rather than lowers, interest rates, which leads to international capital inflows and a currency appreciation. The appreciation decreases net exports and therefore reduces aggregate demand—which shifts inward from D_1D_1 to D_2D_2. But it also reduces foreign prices and raises aggregate supply—as indicated by the shift from S_0S_0 to S_2S_2. The result is that output and prices both fall more in an open economy (point *C*) than they would in a closed economy (point *B*).

the exchange rate, thereby crowding out net exports. But that means that the initial effects of a fiscal expansion on aggregate demand are *weakened* while the initial effects of a monetary contraction are *strengthened*.

INTERNATIONAL ASPECTS OF DEFICIT REDUCTION

This completes our theoretical analysis of the macroeconomics of open economies. Now let us put the theory through its paces by applying it to the prospective changes in U.S. macroeconomic policy over the balance of this decade. Doing so will help us answer the question posed at the beginning of the chapter: Should deficit reduction be expected to strengthen or weaken the dollar?

The multiyear budget program debated in 1995 and 1996 called for fiscal policy to become tighter and tighter over the rest of this century and into the next, as the budget deficit fell from about 2.5 percent of GDP to zero. This movement toward a balanced budget was to have been accomplished mainly by reducing government expenditures.

The first column of Table 36-3 indicates what the theory predicts a fiscal contraction should do: It should lower real interest rates, make the dollar depreciate, reduce real GDP, and be less disinflationary than normal because of the falling dollar. This information is recorded by entering + signs for increases and − signs for decreases.

Now let us consider the monetary policy that may accompany the fiscal contraction. Since reducing the budget deficit reduces aggregate demand, the Federal Reserve would be expected to restore the missing demand by lowering interest rates. According to the analysis in this chapter, such a monetary expansion should lower real interest rates, make the dollar depreciate, raise real GDP, and be a bit more inflationary than usual because of the falling dollar. All this is recorded in Column 2.

TABLE 36-3 EXPECTED EFFECTS OF POLICY

Variable	(1) Fiscal Contraction	(2) Monetary Expansion	(3) Combination
Real interest rate	−	−	−
Exchange rate	−	−	−
Net exports	+	+	+
Real GDP	−	+	?
Inflation	−	+	?

Column 3 puts the two pieces together. We conclude that a policy mix of fiscal contraction and monetary expansion should reduce interest rates strongly, push down the value of the dollar, and act as a strong stimulus to our foreign trade. But its net effects on output and inflation are uncertain; the balance depends on whether the fiscal contraction overwhelms the monetary expansion, or vice versa.

Only time will tell if these predictions come true. But we can "test" the theory by looking back 10 to 15 years to an historical episode in which essentially the same experiment was run in the opposite direction. In broad outline, the early 1980s witnessed a dramatic change in the policy mix toward *tighter* money and much *looser* fiscal policy, as the Federal Reserve's anti-inflation policy collided with the large tax cuts ushered in by President Ronald Reagan. So the predictions for the early 1980s are precisely the opposite of what Table 36-3 shows higher real interest rates, an appreciating dollar, declining net exports, and uncertain effects on real GDP and inflation.

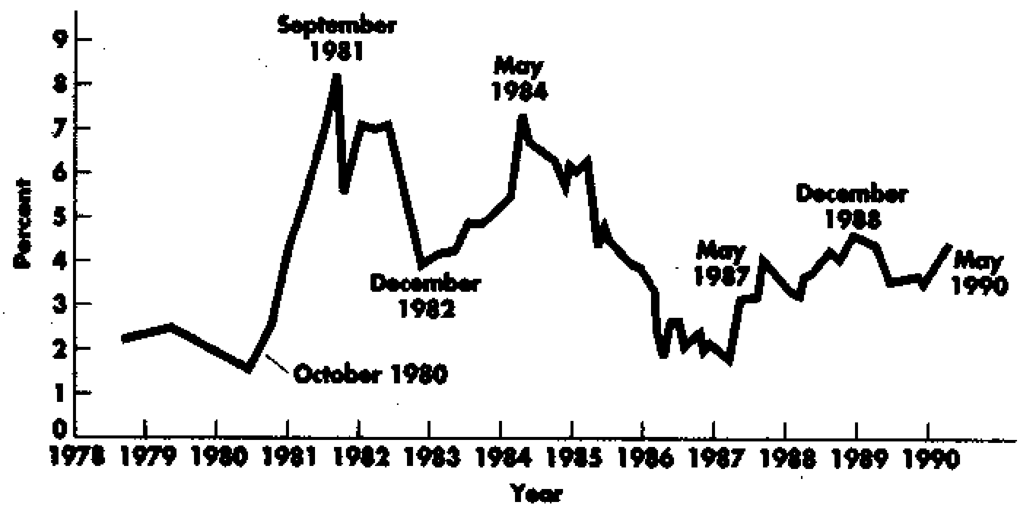
What really happened? We have already observed that the international value of the dollar soared and American foreign trade was clobbered, just as the theory predicts.

But what about interest rates? Figure 36-8 shows an estimate of the real interest rate on long-term U.S. government bonds from 1978 to 1990. (The historic norm for this rate is between 2 and 3 percent.) The real rate of interest rose dramatically between the 1980 election and the time the Reagan economic program was enacted into law (September 1981). It then fell during the 1982 recession, but rose again to very high levels in mid-1984. Again, this is in accord with the theory: Except for a brief recessionary interlude, the policy changes raised real interest rates.

What about real GDP? The U.S. economy suffered through a severe recession in 1981 to 1982 and then rebounded strongly. To appraise the effects of Reaganomics on real output, we must consider a longer period of time. If we look at the period from 1981 to 1989 as a whole, the average annual growth rate of real GDP was 3.2 percent—just a bit below the average growth rate of the previous 20 years. So the conclusion seems to be that the monetary contraction and fiscal expansion combined to have little net effect on the growth rate of aggregate demand.

If this is so, that would leave only the supply shifts caused by the appreciating dollar (see Figures 36-6 and 36-7). The rising dollar certainly helped slow inflation in the early 1980s by holding down the prices of imported inputs. In addition, the fact that the deep recession came early in President Reagan's term of office meant that the economy had a recessionary gap throughout the period

FIGURE 36-8 REAL INTEREST RATES IN THE UNITED STATES, 1978–1990



This figure charts the behavior of the real interest rate on 10-year U.S. government bonds on the basis of a survey of inflationary expectations. It rises steeply in 1980 to 1981.

SOURCE: Richard B. Hoey.

from 1981 to 1986. Finally, as we saw in earlier chapters, energy prices were falling rapidly. Each of these factors played a role in the rapid disinflation.

On balance, then, the theory did quite well in predicting the effects of tight money and rising budget deficits in the 1980s. If it works as well in the future, deficit reduction and looser monetary policy will lead to lower interest rates, a weaker dollar, and strong U.S. trade performance.

THE LINK BETWEEN THE BUDGET DEFICIT AND THE TRADE DEFICIT

Let us now consider the link between the budget deficit and the trade deficit in more detail. To do so, we need only recall two simple pieces of arithmetic.

The first begins with the familiar equilibrium condition for GDP in an open economy:

$$Y = C + I + G + (X - IM)$$

Since GDP can either be spent, saved, or taxed away:[5]

$$Y = C + S + T$$

Equating these two expressions for Y gives:

$$C + I + G + (X - IM) = C + S + T$$

Finally, subtracting C from both sides, and bringing terms involving the government to the left-hand side and terms involving the private sector to the right

[5]If you do not see why, recall that GDP equals disposable income (DI) plus taxes (T), $Y = DI + T$, and that disposable income can either be consumed or saved, $DI = C + S$. These two definitions together imply that $Y = C + S + T$.

leads to an accounting relationship between the budget deficit and the trade deficit:

$$G - T = (S - I) - (X - IM)$$

In words, the government budget deficit, $G - T$, must be equal to the surplus of savings over investment plus the trade deficit, $-(X - IM)$.

This result may seem mechanical and not particularly interesting, but our second piece of arithmetic will bring out both the common sense behind this equation and its importance for understanding recent events.

As we have noted, under floating exchange rates any deficit or surplus in the current account must be balanced by an equal and opposite surplus or deficit in the capital account. Thus, the current account *deficit*, which is $-(X - IM)$ in the previous equation, can be replaced by the capital account *surplus* to get:

$$G - T = (S - I) + \text{Net capital inflows}$$

This last equation makes an obvious point. If the U.S. government runs a budget deficit $(G - T)$, that deficit must be financed either by an excess of saving over investment by American businesses and individuals $(S - I)$, or by borrowing from foreigners.

But what is the economic mechanism that makes the trade deficit follow the government budget deficit? There is no mystery. To attract the necessary foreign capital, America must offer higher interest rates. As capital flows into the United States in response to the higher rates, the value of the dollar rises. And the more costly dollar leads to a larger trade deficit. Thus, the adjustments of interest rates and exchange rates that we have been discussing close the loop between a larger budget deficit and a larger trade deficit.

Applying the same logic in the opposite direction, smaller budget deficits in coming years are expected to lead to lower interest rates, a lower value of the dollar, and smaller trade deficits.

IS THE TRADE DEFICIT A PROBLEM?

The preceding explanation suggests that America's large trade deficits over the last 15 years are a symptom of a deeper trouble: The nation as a whole—including our government—has spent more than it has produced and has been forced to borrow the difference from foreigners. The trade deficit is just the mirror image of the required capital inflows.

Those who worry about trade deficits point out that these capital inflows create debts on which interest and principal payments will have to be made in the future. In this view, Americans have been mortgaging their future to finance higher consumer spending.

But there is another, quite different, interpretation of the trade deficit. Suppose foreign investors in the 1980s began to see the United States as a much more attractive place to invest their capital. Then capital would have flowed here not because Americans needed to borrow it, but because foreigners were eager to lend it. In that case, the trade deficit would still be the mirror image of the capital inflows; however, it would now signify America's economic strength, not its weakness.

The policy implications of these two views differ dramatically. The first view suggests that we should strive to reduce the trade deficit and, thereby, our need for foreign borrowing. The second view suggests that we should welcome the

capital inflows and, therefore, not worry about the trade deficit. Which view is right?

The answer is in dispute. However, most economists in the 1980s ascribed to the first view—that the trade deficit was a problem, not a prize. At least two pieces of evidence persuaded them. First, real interest rates in the 1980s were generally high, not low. This suggested that America was being forced to pay high interest rates to attract the capital it needed, not that foreigners had suddenly decided that America was the place to invest.

Second, as we saw in Table 36-2 (page 854), it was mainly the *consumption* share of GDP, not the *investment* share, that rose between 1981 and 1986. This suggests that America was borrowing to finance a consumption binge, not an investment boom.

ON CURING THE TRADE DEFICIT

We now understand how the United States acquired massive trade and budget deficits in the 1980s and became the world's most indebted nation. And we have explained why many economists think the trade deficit is a problem. Next comes the hard question: What can be done about it? How can we cure our foreign trade problem and end our addiction to foreign borrowing?

The answer, of course, is highly controversial. Both economists and politicians disagree over the appropriate course of action. The best we can do here is outline some alternatives.

CHANGE THE MIX OF FISCAL AND MONETARY POLICY

The fundamental equation:

$$G - T = (S - I) - (X - IM)$$

suggests a *reduction in the budget deficit* as one good way to reduce the trade deficit. According to the analysis in this chapter, a reduction in G or an increase in T would lead to lower real interest rates in the United States, a depreciating dollar, and, eventually, a shrinking trade deficit.

This is the route that American economists urged for years and that the American government has been pursuing in the 1990s. And, as a matter of fact, the dollar has come down considerably since the mid-1980s, and so has our trade deficit (as a share of GDP), although it remains large. Partisan political bickering stood in the way of deficit reduction until President Clinton pushed through a comprehensive deficit-reduction plan in 1993. And the bipartisan support for balancing the budget promises an even tougher fiscal policy.

When the government curtails its spending or raises taxes to reduce its budget deficit, aggregate demand falls. If we do not want deficit reduction to cause an economic contraction in the United States, we must therefore compensate for it by monetary stimulus. Like contractionary fiscal policy, expansionary monetary policy lowers interest rates, depreciates the dollar, and should therefore help reduce the trade deficit.

MORE RAPID ECONOMIC GROWTH ABROAD

If foreign economies grew faster, residents of these countries would buy more American goods. That would raise American exports and reduce our trade

deficit. Since the mid-1980s, the United States has been urging our major trading partners to stimulate their economies and to open their markets more to American goods—with some modest success. In the late 1980s, foreign countries rightly asked why they should tailor their economic policies to America's needs. But since the early 1990s, with the European and Japanese economies often producing well below capacity, domestic needs in those countries have also called for more expansionary policies.

RAISE DOMESTIC SAVINGS OR REDUCE DOMESTIC INVESTMENT

Our fundamental equation calls attention to two other routes to a smaller trade deficit: higher savings or lower investment.

U.S. personal saving rates have been near all-time lows in recent years. If Americans would save more, we could finance more of our government budget deficit at home and therefore would not need to borrow so much abroad. This, too, would lead to a cheaper dollar and a smaller trade deficit. The only trouble is that no one has yet found a reliable way to induce Americans to save more. A variety of tax incentives for saving has been tried, and more are suggested every year. But there is not much evidence that these incentives have worked. We seem to be a nation of consumers.

If the other cures for our trade deficit fail to work in time, the trade deficit may cure itself in a particularly unpleasant way: by dramatically reducing U.S. domestic investment. Let us see how this might work. As our trade deficits and foreign borrowing persist, foreigners wind up holding more and more U.S. dollar assets. At some point, their willingness to acquire yet more dollar assets will begin to wear thin, and they will begin demanding much higher interest rates. At best, higher interest rates lead to lower investment in the United States. At worst, foreigners cease lending to the United States, interest rates skyrocket, and we experience a severe recession. A recession, of course, would reduce our trade deficit substantially by curbing our appetite for imports. But it is a painful cure.

PROTECTIONISM

We have saved the worst remedy for last. One seemingly obvious way to cure our trade deficit is to limit imports by imposing stiff tariffs, strict quotas, and other protectionist devices. We discussed protectionism, and the reasons why almost all economists oppose it, in Chapter 34. Despite the economic arguments against it, protectionism has an undeniable political allure. It seems, superficially, to "save American jobs." And it conveniently shifts the blame for our trade problems onto foreigners.

In addition to depriving us and other countries of the benefits of comparative advantage, protectionism might not even succeed in reducing our trade deficit. One reason is that other nations may retaliate. If we erect trade barriers to reduce our imports, IM will fall. But if foreign countries erect corresponding barriers to our exports, X will fall, too. On balance, our *net* exports, $X - IM$, may or may not improve. However, world trade will surely suffer. This is a game that may have no winners, only losers.

Even if other nations do not retaliate, tariffs and quotas may not improve our trade deficit much. Why? If they succeed in reducing American spending on imports, tariffs and quotas will thereby reduce the supply of dollars on the world

SAVING PATTERNS AND TRADE DEFICITS: THE ARITHMETIC OF U.S.-JAPANESE ECONOMIC RELATIONS

The huge U.S. trade deficit with Japan has been a significant source of friction between the two countries for years and has led to frequent calls for protectionist measures here. Our fundamental equation,

$$G - T = (S - I) - (X - IM)$$

teaches us that part of the problem traces to different saving habits in the two countries. The Japanese are among the biggest savers in the world. So $S - I$ is a large, positive number in Japan. Furthermore, unlike the U.S. government, the Japanese government has often (though not recently) had a budget surplus. It therefore follows that, in order to balance the international books, Japan must generate a large trade *surplus*.

The contrast between the United States and Japan in this regard is marked. While the Japanese people and government together are big net *savers*, the American people and government together are big net *borrowers*. In an integrated world financial system, it is therefore natural that the Japanese should be lending to us. In short, Japan should have capital *outflows* and we should have capital *inflows*—which is just what has been happening. Remembering that:

Current account surplus + Capital account surplus = 0

the implication is that Japan should have a current account *surplus*, and we should have a current account *deficit*.

Once again this is only natural. In fact, the United States has had a trade deficit with Japan for a long time—even when our overall trade position and Japan's were nearly balanced. Being an island nation almost devoid of natural resources, Japan must run huge trade deficits in primary products. Much of this trade is with developing countries. To offset this trade deficit in primary products, Japan needs a surplus in trade in manufactured goods. And who are likely to be the leading customers for these goods? The biggest consumers on earth, of course—the Americans.

So it is natural for the United States to run a bilateral deficit in trading goods with Japan. However, that does not imply that an annual deficit of $50 billion or more is appropriate, nor that the Japanese are blameless. In fact, Japan has long been among the most protectionist of all the advanced, industrial nations. Only a little of this protectionism takes the form of high tariffs or stiff quotas. Most is more subtle, coming instead through a variety of bureaucratic regulations that make importing difficult.

So one possible solution to the U.S.-Japan trade problem is to persuade Japan to open its markets more. And the Clinton administration has devoted considerable time and energy toward this end. But no one really thinks that, even in a completely free market, we could sell in Japan nearly as much as they sell here.

Macroeconomic policy might be a more effective tool. Look once again at the fundamental equation:

$$G - T = (S - I) - (X - IM)$$

If Japan stimulated its economy by more expansionary fiscal policy, $G - T$ would rise and $(X - IM)$ would fall. If, at the same time, the United States reduced its budget deficit, $G - T$ would fall here and $(X - IM)$ would rise. In all likelihood, our bilateral trade deficit with Japan would narrow. That is precisely the "deal" that the two countries are working on now.



market. That would push the value of the dollar up. A rising dollar, of course, would hurt U.S. exports and encourage more imports. The fundamental equation,

$$G - T = (S - I) - (X - IM)$$

reminds us that protectionism can raise $X - IM$ only if it reduces the budget deficit, raises saving, or reduces investment.[6]

CONCLUSION: WE ARE NOT ALONE

We do indeed live in a world economy. The major trading nations of the world are linked by exports and imports, by capital flows, and by exchange rates. What happens to national income, prices, and interest rates in one country affects other nations.

Thus, policymakers in Europe, Asia, and South America keep a watchful eye on developments in the U.S. economy. If the U.S. economy expands, these other countries have better markets for their exports. If we pursue policies that make the dollar depreciate, they find their currencies appreciating. If interest rates rise in the United States, they see capital flowing out of their countries into ours. Some observers think that, as the "big guy on the block," America bears a special responsibility for the health of the world economy.

But we are not the *only* big guys on the block. Japan also has a giant economy which profoundly affects the rest of the world. And, once fully unified, the European Union will give western Europe the biggest economy of all. What happens in Europe, Japan, and elsewhere often has important effects on the U.S. economy.

That the major economies of the world are linked suggests the need for greater policy coordination among nations. But since the national interests of particular countries often differ, countries are understandably reluctant to surrender any of their sovereignty. Hence, international policy coordination remains an elusive goal. Economically speaking, we all live in one world. Politically, however, we live in a world of separate nation-states.

[6]Here tariffs, which raise revenue for the government, have a clear advantage over quotas, which do not.

SUMMARY

1. The nations of the world are linked together economically because national income, prices, and interest rates in one country affect those in another. They are thus **open economies.**
2. Because one country's *imports* are another country's *exports*, rapid (or sluggish) economic growth in one country contributes to rapid (or sluggish) growth in other countries.
3. A country's *net exports* depend on whether its prices are high or low relative to those of other countries. Since exchange rates translate one country's prices into the currencies of other countries, the *exchange rate* is a key determinant of net exports.
4. If the currency depreciates, net exports rise and aggregate demand increases, thereby raising both real GDP and the price level. A depreciating currency also reduces aggregate supply by making imported inputs more costly.
5. If the currency appreciates, net exports fall and aggregate demand, real GDP, and the price level all decrease. But an appreciating currency also increases aggregate supply by making imported inputs cheaper.
6. *International capital flows* respond strongly to interest rate differentials among countries. Hence, higher domestic interest rates lead to currency *appreciations*, and lower interest rates lead to *depreciations*.
7. Contractionary monetary policies raise interest rates and therefore make the currency appreciate. Both the higher interest rates and the stronger currency

reduce aggregate demand. Hence, international capital flows make monetary policy more powerful than it would be in a *closed economy.*

8. Expansionary fiscal policies also raise interest rates and make the currency appreciate. But, in this case, the international repercussions cancel out part of the demand-expanding effects of the policies. Hence, international capital flows make fiscal policy less powerful than it would be in a closed economy.
9. Since Reaganomics in the early 1980s combined tight money with highly expansionary fiscal policy, it raised interest rates, pushed the dollar up, and caused a large trade deficit in the United States. Conversely, deficit reduction and easier money in the 1990s are expected to reduce interest rates, the value of the dollar, and the trade deficits.
10. *Budget deficits and trade deficits* are linked by the fundamental equation $G - T = (S - I) - (X - IM)$. This also implies that the budget deficit equals the sum of $S - I$ + *Capital inflows.*
11. It follows from this equation that the U.S. trade deficit must be cured by some combination of lower budget deficits, higher savings, and lower investment.
12. Protectionist policies might not cure the U.S. trade deficit because (a) they will make the dollar appreciate and (b) they may provoke foreign retaliation.

KEY TERMS

Exports
Imports
Net exports
Closed economy
Open economy
Exchange rate
Appreciation
Depreciation
Trade deficit
International capital flows
Budget deficits and trade deficits
$G - T = (S - I) - (X - IM)$

QUESTIONS FOR REVIEW

1. For years, the U.S. government has been trying to get Japan to expand its economy faster. Explain how more rapid growth in Japan would affect the U.S. economy.
2. If inflation is higher in Germany than in France, and the exchange rate between the two countries is fixed, what is likely to happen to the balance of trade between the two countries?
3. Explain why a currency depreciation leads to an improvement in a country's trade balance.
4. Explain why American fiscal policy is less powerful and American monetary policy is more powerful than indicated in the closed-economy model of the previous part of this book.
5. Use an aggregate supply-demand diagram to analyze the effects of a currency appreciation.
6. Explain why $G - T = (S - I) - (X - IM)$.
7. Given what you now know, do you think it was a good idea for the United States to adopt a policy mix of tight money and large government budget deficits in the early 1980s? Why or why not?
8. What, in your view, is the best way for America to reduce its trade deficit?
9. During 1994 and early 1995, the international value of the yen rose sharply. This development worried the Japanese authorities. Why?
10. (More difficult) Suppose consumption and investment are described by:

$$C = 50 + 0.75DI$$
$$I = 400 + 0.2Y - 50r$$

Here DI is disposable income, Y is GDP, and r, the interest rate, is measured in percentage points. (For example, a 5 percent interest rate is $r = 5$.) Exports and imports are as follows:

$$X = 200$$
$$IM = -150 + 0.2Y$$

Government purchases are $G = 800$, and taxes are 20 percent of income.

The price level is fixed and the central bank uses its monetary policy to peg the interest rate at $r = 8$.

a. Find equilibrium GDP, the budget deficit or surplus, and the trade deficit or surplus.
b. Suppose the currency depreciates and, as a result, exports and imports change to:

$$X = 250$$
$$IM = -200 + 0.2Y$$

Now find equilibrium GDP, the budget deficit or surplus, and the trade deficit or surplus.

CHAPTER 37

PRODUCTIVITY AND GROWTH IN THE WEALTH OF NATIONS

The enormous wealth of Britain is attributable [primarily] to . . . the wonderful skill of her entrepreneurs . . . and the superiority of her workmen in rapid and masterly execution. . . . The English laborer . . . gives his work more care, attention and diligence, than the workmen of most other nations.
Jean Baptiste Say, 1819

Japan commercially, I regret to say, does not bear the best reputation for executing business. Inferior goods, irregularity and indifferent shipments have caused no end of worry . . . you are a very satisfied easy-going race . . . the habits of national heritage.
From a report of an Australian expert for the Japanese government, 1915

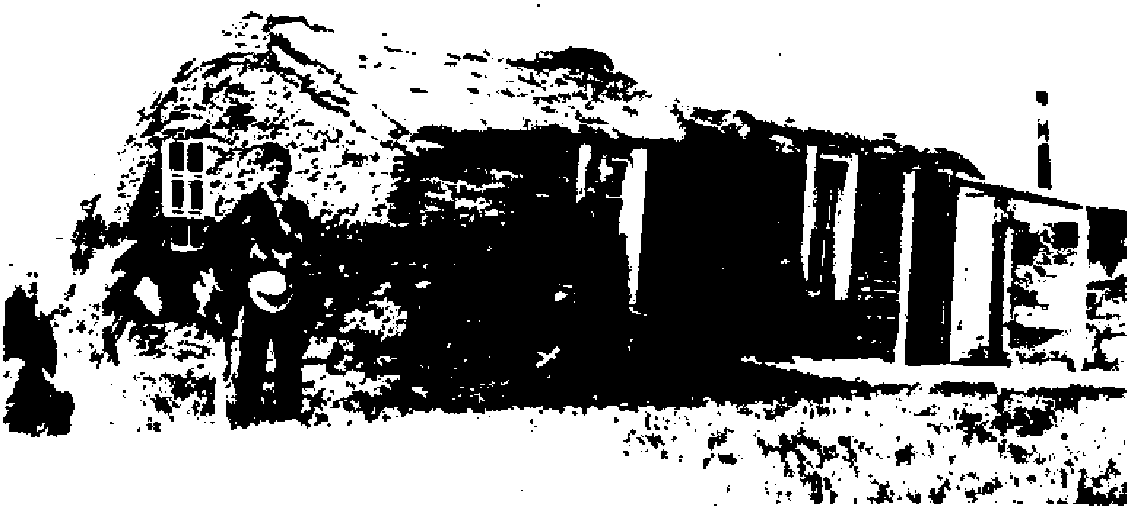
This chapter introduces a major topic of growing importance for all Americans: the role of productivity in the well-being of nations. We are conscious as never before of competition from other industrial economies. Our prosperity depends increasingly on that of other countries, and theirs, in turn, depends vitally on ours. This chapter compares for a number of countries the growth in living standards and *productivity* (output per hour of work) over many decades. The chapter is divided into two parts. The first focuses on the wealthier, industrialized nations and the second deals with the poorer, less developed countries (LDCs).

PART 1: GROWTH IN THE UNITED STATES AND OTHER INDUSTRIAL COUNTRIES

Human history in the last two centuries has been unlike anything ever experienced before. In the world's industrial countries, the quantity and quality of food, clothing, and comforts have reached levels that were never before dreamed possible. The change has been so revolutionary that it is difficult to grasp. This chapter helps us envision how great the accomplishment has been. It also suggests that the transformation was made possible by productivity growth: the fact that the hourly production of a person in the United States today is perhaps 20 times as high as it was in 1800. Just two figures will suggest the magnitude of the achievement. In 1800 about 90 percent of America's labor force had to work on farms, but all that farm labor barely managed to produce enough food to feed the country adequately. Today less than 3 percent of U.S. workers earn their livings on farms. Yet those few farm workers provide an outpouring of surpluses which the U.S. government constantly struggles to contain.

LIFE IN THE "GOOD OLD DAYS"

The United States, and the 13 colonies before it, has always been a privileged land with relatively high levels of nutrition. In the 18th century, an average white, native-born male who reached the age of 10 could expect to live to



Late 19th-century sod house in the Dakotas.

somewhere between age 50 and 55. By contrast, an English *nobleman* at that time could expect to live only to something between age 39 and 46.

Nevertheless, in the mid-19th century, low farm productivity, an almost complete lack of refrigeration, and limited transport of goods bound a large part of even the U.S. population to a minimal and nutritionally inferior variety of foods. Such uninspiring staples as lard, cornmeal, and salt pork were mainstays in the diet, particularly outside the population centers. Travelers' accounts of meals often mentioned the universal one-pot stew, the main meal of the day for the family. According to one study, "There were, of course, a few people who knew what it was to . . . eat a meal that consisted of more than one course; but there were very, very few such people, and they were all very rich."[1]

But, most Americans rightly felt they lived in a land of unprecedented abundance, for that one-pot stew was usually there. For many centuries most Europeans had devoted nearly half their food budgets to breadstuffs. For example, in 1790 in France, "The price of bread [for a family of five], even in normal times . . . was half . . . the daily wage of common labor."[2] Often the bread took the form of gruel—a sort of cooked breakfast cereal—served in a single bowl with a single spoon, both passed around the table to the entire family. In bad years, even gruel was unavailable. Famine and widespread death by starvation continued to threaten Europe until the beginning of the 19th century, and earlier it had constituted a normal fact of existence. Food shortages were but one manifestation of the unimaginably poor living conditions in all the major cities of Europe.

Even in 19th-century America, though, living conditions were far from ideal. Most rural homes that travelers saw were tiny and crudely built, with no glass windows, no lighting except the fireplace, no indoor plumbing, and scanty homemade furniture. The housing of relatively well-off Americans was still primitive by modern standards. Bathtubs, for example, were rare. No homes

[1]Ruth Schwartz Cowan, *More Work for Mother: The Ironies of Household Technology from the Open Hearth to the Microwave* (New York: Basic Books, 1983), p. 38.

[2]Robert Palmer, *The Age of Democratic Revolution*, vol. 2 (Princeton, N.J.: Princeton University Press, 1964), p. 49.

had electricity and few had gas. Fewer still had hot running water, and not even 2 percent had indoor toilets and cold running water. Boston, with a population of nearly 200,000 in 1860, had only 31,000 sinks, 4,000 bathtubs, and 10,000 water closets. Outdoor privies were the norm and baths, for the great majority, a luxury.

Like the common man, the rich have also gained much in terms of health and personal comfort in the course of two or three centuries. By the early 1900s, life expectancy at birth for a member of the British nobility had reached 65 years. Home heating technology illustrates the dramatic improvements in "creature comforts." It was not until early in the 18th century that heating technology made any improvements on the open fireplace. Until then the bitter cold of winter was a serious threat to both the poor and the highest nobility. Princess Palatine, the German sister-in-law of King Louis XIV, reported that in February 1695, "in the Hall of Mirrors at Versailles at the King's table the wine and water froze in the glasses."[3]

Though living conditions were vastly improved from earlier centuries, by today's standards, life in the United States just 100 years ago was hard and primitive.

THE MAGNITUDE OF PRODUCTIVITY GROWTH

Today, of course, things are vastly different. Recent surveys indicate that only 1.1 percent of American housing units lack complete plumbing—defined as hot and cold piped water, a flush toilet, and a bathtub or shower for the exclusive use of that housing unit. Less than 5 percent are occupied by more than one person per room. Of the new, privately owned, one-family houses built in 1993, 58 percent had three or more bedrooms, 49 percent had 2.5 or more bathrooms, and 79 percent had central air conditioning. Furthermore, in that year over 98 percent of all households owned a television (with an average of 2.2 sets per home), virtually 100 percent were equipped with electric refrigerators, 76 percent owned electric washing machines, 79 percent owned microwave ovens, and 55 percent had two or more vehicles.[4]

The revolution in manner of living was made possible by an unprecedented rate of growth in human efficiency in producing output.

LABOR PRODUCTIVITY refers to the amount of output a worker turns out in an hour (or a week or a year) of labor. It can be measured as total national output (GDP) in a given year divided by the total number of hours of work performed for pay in the country during that year. That is, labor productivity is defined as GDP per hour of labor.

PRODUCTIVITY AND OUTPUT PER CAPITA

There are two basic concepts for the long-run growth story: *labor productivity* and *output per capita (per person)*. They are, respectively, the indicators of the production and the consumption side of economic growth.

Labor productivity refers to the amount of output turned out by a *given* amount of labor [*gross domestic product (GDP) per labor hour*]. An increase in productivity means that a human being has become a more effective instrument of production. This can be the result of harder work, better training, more or better equipment, innovative technology, or a variety of other causes.

[3]Fernand Braudel, *The Structures of Everyday Life: The Limits of the Possible*, vol. 1, *Civilization and Capitalism, 15th–18th Century* (New York: Harper & Row, 1979).

[4]U.S. Bureau of the Census, *Statistical Abstract of the United States, 1995*, 115th ed. (Washington, D.C.: U.S. Government Printing Office, 1995).

The *standard of living*, on the other hand, is more naturally measured by *GDP per capita*, that is, by total output (GDP) divided by the number of persons among whom it will be distributed. The more output there is for each person, the better off in economic terms the average person must be.

We usually measure a nation's prosperity not in terms of its total output but in terms of its output *per person*. For instance, India has a GDP more than seven times as large as Sweden's. But with a population more than 100 times as large as Sweden's, India remains a poor country while Sweden is highly prosperous. The point is that:

If the objective of growth is the material welfare of the *individuals* who make up a country, then the proper measure of the success of a program of economic development is how much it adds to output per person—total output *divided by total population*.

To make the appropriate comparison of well-being in Sweden and India, we note that per-capita GDP in Sweden is over $18,000 a year, whereas in India, even after a generous adjustment to correct for lower prices in that country, the figure is about $1,360 a year.[5]

The fantastic magnitude of the increases in both labor productivity and output per capita since 1800 is best appreciated by contrasting it with the dismal average record of many previous centuries. It is estimated that in Europe, even by the time of the American Civil War, neither labor productivity nor GDP per capita had yet come back up to the levels that had been achieved in ancient Rome! Thus, on the average, *productivity and GDP per capita did not grow at all for over 1,600 years!* Even for those wealthy enough to buy them, the number of important new consumer goods introduced in all those centuries was remarkably small. Firearms, glass windowpanes, eyeglasses, mechanical clocks, tobacco, and printed books constitute almost the entire list of major new consumer products invented between the fall of the Roman Empire and the beginning of the 19th century.

In contrast, the period since the 1830s has been characterized by an endless explosion of innovations. The railroad and the steamship revolutionized transportation. Steel-making technology changed drastically. The chemical and electronics industries were born and produced hundreds of new consumer products. The range of personal and household goods that we now take for granted—TV sets, dishwashers, cameras, automobiles, personal computers, and many, many others—appeared in an accelerating stream and became commonplace. This has reached a point where today our one unchanging expectation for the future is that it will be characterized by constant change.

Figure 37-1 shows, for five countries, the impressive growth of labor productivity over the past century.[6] For example, it indicates that between 1870 and

---

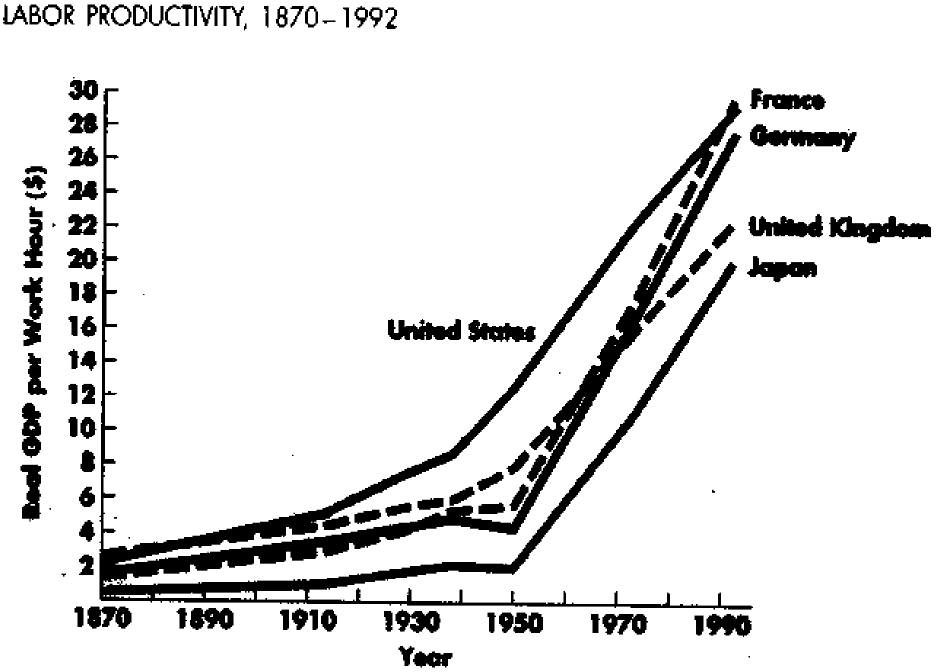
[5]These comparisons are based on per-capita GDP figures, expressed in U.S. dollars and converted from the national currencies using purchasing-power-parity exchange rates. The source is the Central Intelligence Agency, *The World Factbook 1995* (Washington, D.C.: CIA, 1995). For an explanation of purchasing-power-parity rates, see footnote 6 below.

[6]Throughout this chapter, as we compare output per capita or output per work hour (productivity) in different countries, we run into a problem. U.S. output is measured in dollars and other countries' outputs in their own currencies. It is *not* legitimate to compare them by finding out the amounts of different foreign currencies that your bank will give you for a dollar, because those exchange rates change from day to day, but relative productivity levels do not. Instead we try to measure the outputs of different countries in money (usually dollars) of constant purchasing power. For example,

FIGURE 37-1 LABOR PRODUCTIVITY, 1870–1992

The productivity growth shown here for five industrial countries is typical of today's leading industrial economies. The explosive pattern of increase is unprecedented.

SOURCE: Angus Maddison, *Monitoring the World Economy, 1820–1992* (Paris: Development Centre of the Organisation for Economic Cooperation and Development, 1995).



1992 Japanese productivity has risen about 4,300 percent; French and German productivity levels went up 2,100 percent and 1,600 percent, respectively; productivity in the United States increased about 1,200 percent; and British productivity jumped 800 percent!

To see the implications of this dramatic rise in productivity for living standards (as measured by output per capita) we must first note what has happened to labor time spent by the typical worker. In the industrialized, free-market countries, the average number of hours worked per year has fallen by half. This is partly the result of a fall in work hours per day—typically from about 12 hours in 1900 down to some 7.5 hours per day by 1979—and partly the result of a decline from six working days to five. But most surprising is the almost total absence of any vacations for most of the population before the 20th century. The two-week or four-week vacation is another luxury made possible by the rise in productivity.

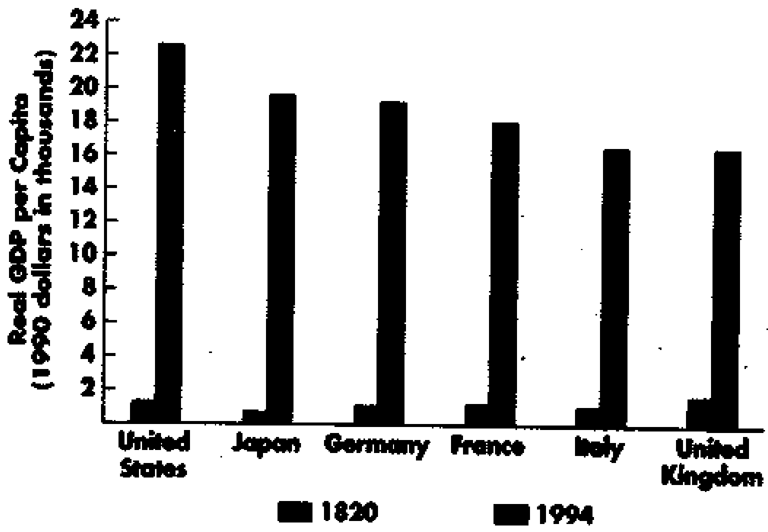
Increases in output per person (or average income per head) have also been spectacular. Over the 174-year period from 1820 to 1994, output per capita went up almost 2,700 percent in Japan, about 1,600 percent in Germany and the United States, about 1,400 percent in Italy and France, and more than 800 percent in Great Britain.[7] In the United States, for example, average income measured in

---

we ask how many yen does it cost to buy the same bundle of goods in Japan that one can buy for, say, $100 in the United States. This figure is referred to as a *purchasing-power-parity* exchange rate. Such numbers are used in all the international comparisons in this chapter.

[7] It is important to notice that all of these numbers have been corrected to eliminate the effects of inflation. Thus, it is true, of course, that because of the subsequent inflation, a dollar in the mid-1800s could purchase much more bread or many more shoes than it can today, so that a $12 weekly salary then is not as low, in purchasing power, as it may seem. But the statistics reported here have been corrected to eliminate this source of confusion, using GDP price indexes. See the appendix to Chapter 23 for further explantion.

FIGURE 37-2 IMPROVEMENT IN LIVING STANDARDS, 1820 VERSUS 1994



The graph shows how much higher GDP per capita is today than it was in 1820 in each of the six countries shown. The pattern is typical for free-market, industrialized countries. The numbers are in dollars adjusted to have roughly the same purchasing power in all countries and at both dates listed.

SOURCE: Angus Maddison, *Monitoring the World Economy, 1820–1992* (Paris: Development Centre of the Organisation for Economic Co-operation and Development, 1995).

dollars of constant purchasing power was only about one-seventeenth as large in 1820 as it is today.

To imagine what it is like to live on an income so small, one must think of *Egypt, Bolivia, or the Philippines,* whose per-capita incomes today have been calculated to be on a par with that of an average American in the 19th century! Figure 37-2 translates the productivity growth since the beginning of the 20th century into the resulting and extraordinary rise in living standards—in national output per capita.

Obviously, the economic well-being of an average American, and of an average resident of the other countries in Figures 37-1 and 37-2, has increased dramatically. *It is noteworthy that, according to the World Bank's latest estimates, though other countries are catching up, the United States still leads the world in per-capita GDP.*

Since 1800, productivity in the United States has increased at an average annual rate of slightly less than 2 percent. An apparently small change in this figure would have enormous consequences over a long period. Had productivity grown at an average rate of only 1 percent per year instead, an average American today would command about 6.6 times as large a quantity of goods and services as his forebears did in 1800. Actually, though it is hard to believe, real per-capita income has risen about 30-fold in this period. And if productivity had grown over the entire interval at an annual rate of 3 percent, average living standards would be an incredible 275 times as high as they were in 1800.

Productivity growth can make an enormous difference in a nation's standing in the hierarchy of the world's economies. Keeping its annual productivity growth just one-half of 1 percent ahead of Great Britain for about a century transformed the United States from a minor, developing country into a superpower, while Great Britain declined from the world's preeminent power into a second-rate economy. Japan's 3 percent average annual productivity growth rate since 1870 also transformed it from one of the world's poorest countries into a nation with one of the highest GDP figures in the world.

THE OVERWHELMING IMPORTANCE OF PRODUCTIVITY GROWTH IN THE LONG RUN

As we pointed out in our list of **Ideas for Beyond the Final Exam,** only rising productivity can raise standards of living in the long run.

Over long periods of time, small differences in rates of productivity growth compound like interest in a bank account and can make an enormous difference to a society's prosperity. Nothing contributes more to reduction of poverty, to increases in leisure, and to the country's ability to finance education, public health, environmental improvement, and the arts.

THE SECOND MAJOR DEVELOPMENT: CONVERGENCE

Not only has each of the industrial countries grown in productivity and income per capita, but these countries have also become more similar to one another both in terms of labor productivity and GDP per capita. In other words, among the industrial countries, those which were furthest behind in 1870 have been catching up with those that were ahead. For example, in 1870 the productivity level of the leading country (Australia) was about seven times as high as that of the least productive country (Japan). In 1992 that ratio was only about 1.5. That is, about three-quarters of the difference between the most productive and the least productive countries was eliminated during the 122-year period.

Such a dramatic narrowing of productivity gaps among the industrial countries means that everyone else must be catching up with the leader, and that leader, ever since World War I, has been the United States. This is illustrated in Figure 37-3, which shows what happened to the *relative* labor productivity levels of four leading industrialized countries over the period 1950 to 1992. Specifically, it shows GDP per work-hour in each of those countries *as a percentage of the U.S. level.* (The U.S. figure is, therefore, always 100 percent.) It indicates that, in 1950, per-capita GDP for the average of the countries shown was about 28 percent of that of the United States. By 1992 the average was about 89 percent of the U.S. level. Figure 37-3 also shows that productivity in *each* of these countries has been moving closer to the United States in this 42-year period, with France caught up and Germany not far behind.[8]

These and other data indicate that levels of labor productivity are converging among the leading industrialized countries.[9]

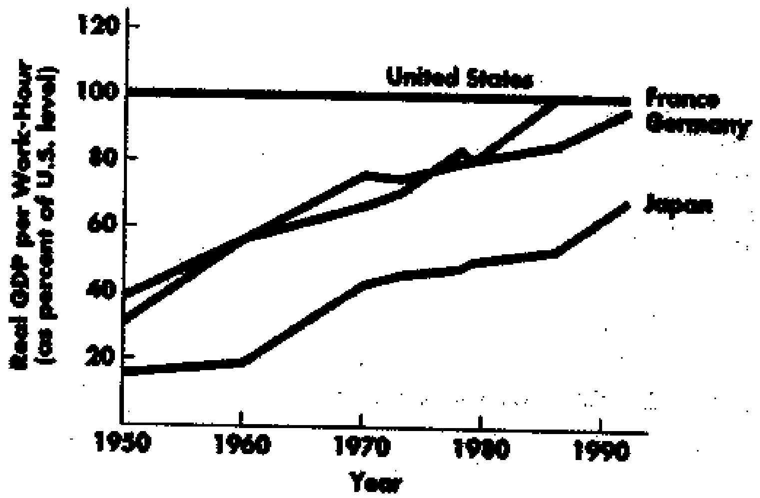
Figure 37-3 also indicates that, as productivity levels in the other countries have come closer to those in the United States, the speed at which they have approached has slowed. Note also that, despite the impression given in the press, Japan is still the lowest in this group, and that Germany is falling behind France (and several other countries).

---

[8]To interpret the graph, remember that if the curve representing some country reaches precisely the curve for the United States, it means that this country's productivity per worker is equal to that of the United States.

[9]Note that this conclusion has been challenged, at least for the years before World War II, on the grounds that the sample of countries studied happens to include those that have been converging toward the United States because they are the success stories for which statistics are available. Thus, the critics point out, Argentina is omitted from the sample of countries, even though in 1870 many observers would have predicted a brilliant economic future for that country.

FIGURE 37-3 LABOR PRODUCTIVITY LEVELS AS A PERCENT OF U.S. LEVEL, 1950–1992



The graph shows that real GDP per worker in each of the countries approached the (growing) U.S. level during the period 1950 to 1992. However, as productivity in those countries got closer to our own, the catch-up process seemed to slow. You may also be surprised to see that in 1992 Japan was still at the bottom of the group and that France was ahead of Germany (and has caught up to the U.S. level).

SOURCE: Angus Maddison, *Monitoring the World Economy, 1820–1992* (Paris: Development Centre of the Organization for Economic Cooperation and Development, 1995).

WHY INTERNATIONAL EQUALIZATION?

Why are these countries growing more alike in productivity and average standards of living? No one has the entire answer, but a good part of the story is probably the speed with which new technology spreads around the world. Better communications permit innovative techniques to move from one country to another far more quickly than in the past. Better and more widespread education permits countries to learn technical details from one another and to train their labor forces rapidly to make use of them. At the beginning of the 18th century when the Newcomen steam engine (the predecessor of Watt's steam engine) was invented in England, it took half a century for the engine to spread to western European countries and to the American colonies. In contrast, the innovations in transistor and semiconductor technology since World War II have, on average, taken about two and one-half years to spread among countries.

All industrial countries benefit from the process of shared information; each learns from the innovations that occur in all the others. The British, the French, and the Germans benefit from American computer technology while the United States and others benefit from Japanese advances in robotics.

But there is one crucial asymmetry. Countries that are behind have a great deal to learn from countries that are ahead, but the leaders have less to learn from those that have lagged behind. This is generally believed to be a prime explanation of the convergence in living standards.

Lagging countries have more to learn from leaders than leaders can learn from laggards. This fact, and the growing speed with which innovations are spread, help explain why the world's industrial economies are growing more equal.

ARE ALL COUNTRIES PARTICIPATING IN EQUALIZATION?

So far we have seen that the world's industrialized countries are growing increasingly similar to one another in terms of productivity and living standards. But we have yet to consider the world's poorest nations—usually referred to as *less developed countries, or LDCs*. (See Part 2 of this chapter.) Have most of those

countries also benefited from the spread of innovation and closed the gap between themselves and the world's economic leaders?

Unfortunately, among the LDCs, where equalization is most desperately needed, the picture is very mixed. Some countries, such as Taiwan and South Korea, have achieved spectacular successes. But, as a group, the LDCs have grown less equal among themselves and have fallen further behind the United States. On average, GDP per capita grew about 3 percent per year in the industrialized countries in the period after World War II. But in the LDCs it rose only about 1 and 1/2 percent a year on average.

While the leading economies in the world are becoming more equal in terms of productivity and living standards, a number of the poorest countries are falling further behind and are holding back the average performance of the LDCs. Later in the chapter we will discuss the reasons cited for the poor performance of many LDCs.

THE U.S. PRODUCTIVITY SLOWDOWN: IS AMERICAN ECONOMIC LEADERSHIP DOOMED?

Since the mid-1960s there has been a sharp decline in the growth rate of productivity in the United States. From 1950 to 1973, productivity per work-hour grew at an estimated average rate of 2.7 percent per year, which is probably somewhat faster than it had ever grown before over so long a period. Then, between 1973 and 1992 productivity growth fell sharply from its earlier postwar rate, back 1.1 percent per year.

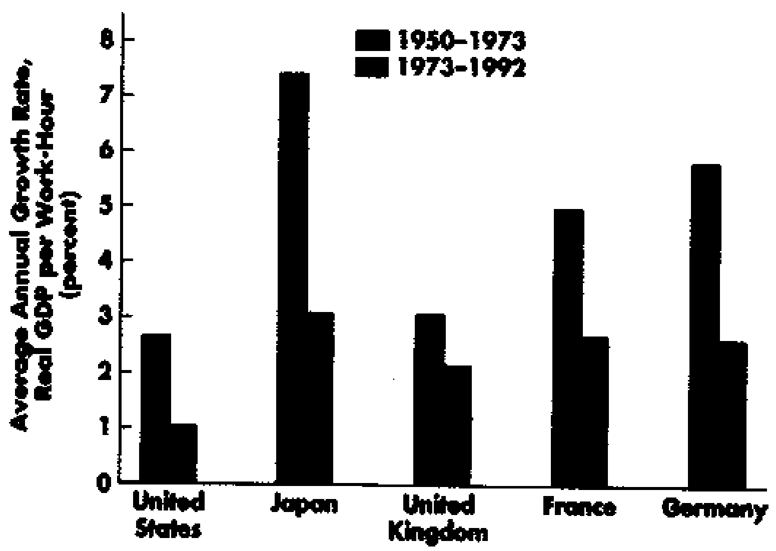
Some observers have concluded that the U.S. economy, and particularly its manufacturing sector, are in terrible trouble and that the United States may lose its leadership in both productivity and living standards. But these fears are, at the very least, exaggerated for several reasons.

First, it is important to note that throughout most of this period the *level* of U.S. productivity continued to improve. In almost every year it was higher than the last. The *rate of improvement* simply slowed from a gallop to a walk, and finally to a crawl. Second, the United States is still by far the world's largest economy in terms of total production, wealth, and factors such as spending on research and development.

Third, some analysts question whether the U.S. productivity slowdown is as serious as it appears. They argue that the United States is not alone in suffering a slowdown in productivity growth. The problem has affected virtually every industrial country. Figure 37-4 compares productivity growth rates for five industrial countries from 1950 to 1973 with those from 1973 to 1992—the years when the slowdown occurred. We find that the U.S. growth rate fell 58 percent, but the growth rates of Japan, the United Kingdom, France, and Germany also fell nearly as much. Compared with the others, the U.S. decline does not seem so far out of line.

In addition, very-long-run data on U.S. productivity growth exhibit no long-term downward trend. Productivity plunged during the Great Depression of the 1930s then leaped upward with the end of World War II. But this catch-up period also came to an end, and the growth rate fell from its postwar high. From this view, then, the deceleration after the 1970s, rather than being entirely a drop below its historical norms, was partly a return toward normalcy from a period of extraordinary growth, which (in retrospect!) seems to have been predictably

FIGURE 37-4 FALL IN PRODUCTIVITY GROWTH RATES, 1950–1973 VERSUS 1973–1992



This graph confirms that productivity growth *did* decline in the period widely publicized as the "U.S. productivity slowdown." However, at the same time, it also fell throughout the industrial world and not just in the United States.

SOURCE: Angus Maddison, *Monitoring the World Economy, 1820–1992* (Paris: Development Centre of the Organization for Economic Cooperation and Development, 1995).

temporary. The data do indicate that since 1973 the growth rate of the U.S. productivity (an average annual rate of 1.1 percent) continues below the historical figure (about 2 percent). However, per-capita income has done better than productivity. The average annual growth rate of U.S. GDP per capita between 1954 and 1977 (2.1 percent) was, indeed, significantly higher than it has been since (1.6 percent).[10] However, this more recent figure is *exactly* the same as the average growth rate in the earlier 84-year period between 1869 and 1953 (again, 1.6 percent).

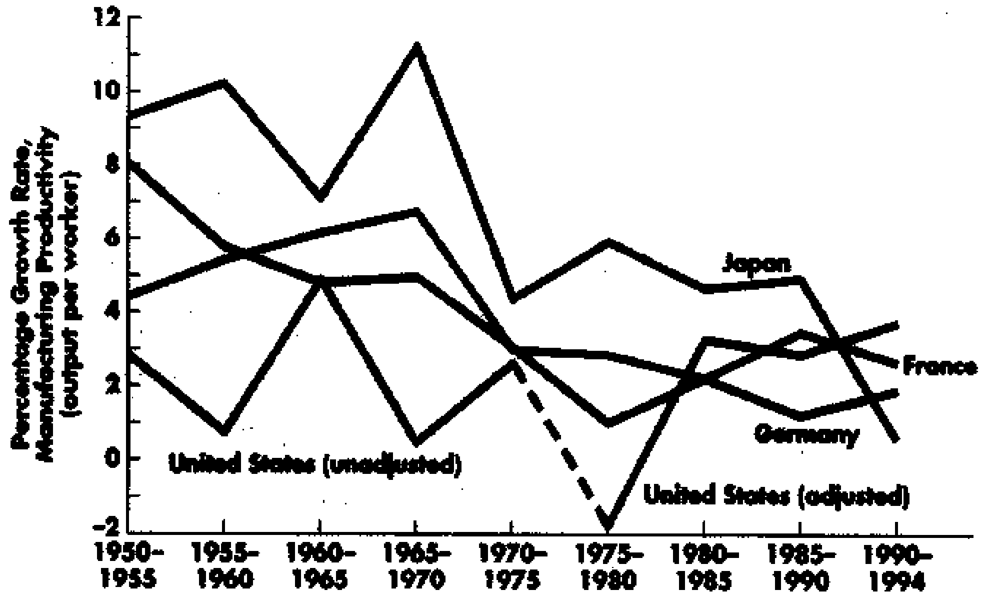
During the recent decline in the growth rate of *overall* U.S. productivity, some sectors—such as construction, mining, and the service industry—have had poor recent records. However, the main focus of public concern has been *manufacturing;* Americans seem to believe that this sector of our economy is most vulnerable to foreign competition because of its poor productivity performance. Yet the record shows something quite different: *U.S. manufacturing productivity has experienced no trend toward a declining growth rate.* This is shown in Figure 37-5 which reports the growth rate of labor productivity in American manufacturing over the entire period since World War II. No downward trend is apparent. Moreover, we see that the growth rates of manufacturing productivity in Japan and Germany are no longer far out of line with those of the United States.

THE U.S. PRODUCTIVITY GROWTH LAG BEHIND OTHER INDUSTRIAL COUNTRIES

What about the second disturbing fact about U.S. productivity growth—that it has recently been so much lower than that of other industrial countries? There is a simple and plausible explanation. It is the equalization of productivity in the world's industrial countries—the convergence phenomenon we just studied. If the forces making for equalization did in fact dominate the growth paths of those countries, it is necessarily true that productivity in the lagging countries *had* to

[10] "What's Happening to Americans' Income?" *The Southwest Economy*, Federal Reserve Bank of Dallas, Issue 2 (1995), pp. 3–6.

FIGURE 37-5 GROWTH RATES IN MANUFACTURING PRODUCTIVITY, 1950-1994



In the important manufacturing sector of the economy, the graph shows that U.S. productivity has shown no sign of a declining trend. Moreover, the growth rates in other leading economies are now coming closer to the American level.

Note: U.S. data for 1950–1976 are based on 1982 price weights; data for 1977–1994 are based on 1987 price weights.

SOURCE: U.S. Department of Labor, Bureau of Labor Statistics, *BLS News: International Comparisons of Manufacturing Productivity and Unit Labor Cost Trends, 1994* (Washington, D.C.: U.S. Department of Labor, September 8, 1995); and various BLS news releases.

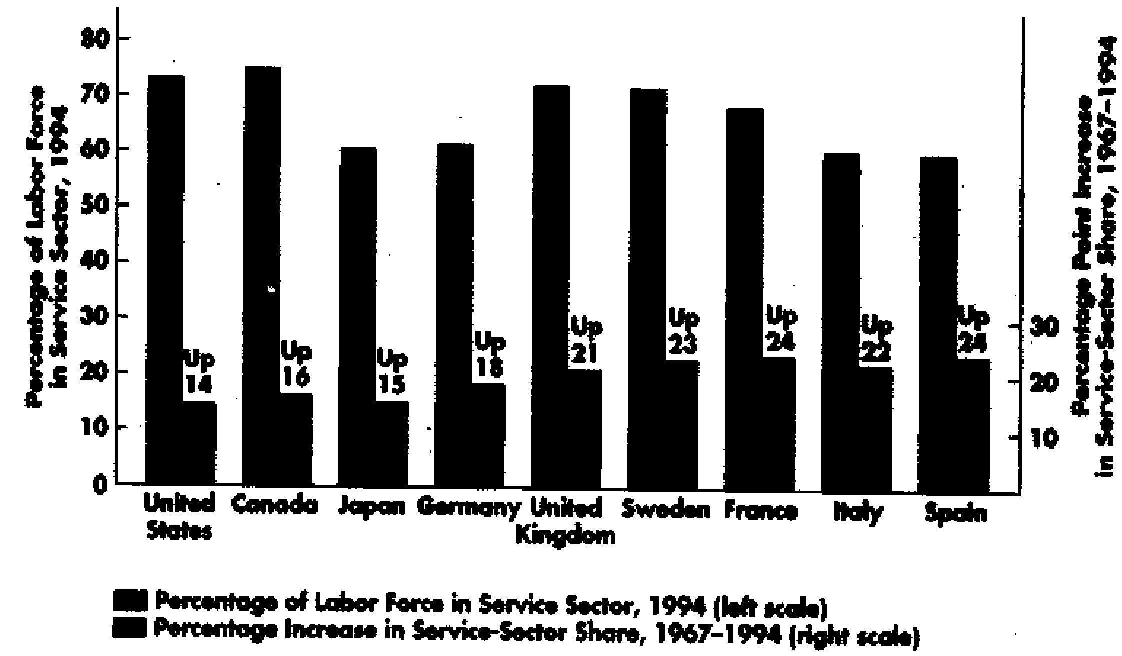
grow more quickly than that in the countries at the head of the line. Otherwise, they could never have grown more equal.

Viewed in this way, there is little to be alarmed about in the relatively slow growth rate of U.S. productivity compared with that of other countries. Indeed, if the higher growth rates of other countries are attributable in good part to their having much to copy from the United States as a productivity leader, we should not be surprised if the rapid growth rates of those countries were to slow as they approach the high American levels. There are, as a matter of fact, signs that this is beginning to happen.

PRODUCTIVITY AND THE DEINDUSTRIALIZATION THESIS

Let us now turn to the popular notion that lagging productivity growth is turning the United States into a service economy, the "deindustrialization thesis." The trends are said to portend a future in which the United States suffers chronic and apparently incurable problems in its trade with other countries because its manufactured products are not competitive with those of foreign countries. As a result, it is argued, the United States will either be forced to bear heavy unemployment or to see its labor force driven into low-paying, service-sector jobs, thus transforming the nation into a "service economy" in which people earn their livings by flipping hamburgers and washing dishes. The "deindustrialization"

FIGURE 37-6 GROWTH IN THE SHARE OF SERVICE-SECTOR JOBS, 1967–1994



The brown bars show how much of the labor force works in the service sector in the corresponding industrial countries. The blue bars indicate how much shares have increased in percentage points since 1967. The graph demonstrates that the United States is hardly the only country in which the proportion of labor employed in the services has been rising. Indeed, the shares of service-sector jobs have grown far more rapidly in many other countries.

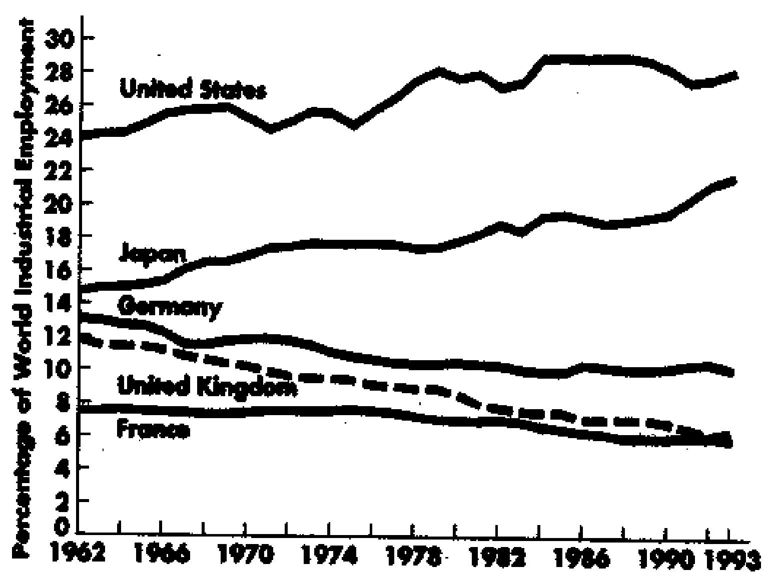
SOURCE: Organization for Economic Cooperation and Development, *Labour Force Statistics, 1970–1990*, Paris: OECD, 1992, and *Quarterly Labour Force Statistics*, Number 1, 1995.

story, oversimplified, asserts that slow productivity growth in manufacturing allows other countries to steal our industrial markets away.

At first glance, the data seem to confirm this. Between 1967 and 1994 the share of the U.S. labor force engaged in industry fell about 12 percentage points, while that in the services rose by about 14 percentage points, just as the deindustrialization thesis asserts. (The difference between these two numbers is accounted for by a 2-percentage-point fall in the share of labor in agriculture.) But, as shown by the blue bars in Figure 37-6, the story breaks down when we seek to identify the countries that have supposedly stolen our industrial markets. The data, for nine leading industrial countries, show that *every* country in the sample has increased the share of its labor force in the services *by a greater percentage than ours*. Which country, then, was "industrialized" by the "deindustrialization" of the United States? Or are all industrial nations becoming service economies and, if so, why?

It turns out that there is a straightforward answer in which productivity plays a key role; but it is very different from the deindustrialization parable. The simple explanation is that throughout the industrial world productivity in manufacturing has grown considerably faster than it has in most services. For example, productivity has grown far faster in automobile manufacturing than in selling real estate. This means that, though manufacturing outputs have grown, less and less of each nation's labor force has been needed to produce them, just as had happened previously in agriculture. So a declining share of each nation's

FIGURE 37-7 SHARES OF WORLD INDUSTRIAL EMPLOYMENT, 1962–1993



The graph shows that over the 31 years for which the data are available, Japan has steadily gained a larger share of the world's industrial jobs. But the U.S. share also rose significantly, at the expense of France, Germany, and the United Kingdom.

Note: World industrial employment here encompasses the 25 member countries of the OECD, which includes the bulk of the free-market, industrial economies of the world.

SOURCE: Organization for Economic Cooperation and Development, *Indicators of Industrial Activity*, various issues; and *Labour Force Statistics*, various issues.

jobs has been provided in the manufacturing sector just as, in earlier decades, a falling portion of the nation's jobs was provided by agriculture.

Thus, the growing share of service workers in U.S. employment is not attributable to lack of competitiveness of U.S. manufacturers. On the contrary, between 1962 and 1993 the U.S. share of the total industrial employment of the world's 25 most industrialized economies actually *increased* about 4 percentage points. (See Figure 37-7, which shows the nearly parallel growth of the U.S. and Japanese shares of world industrial jobs.) That is hardly a picture of faltering competitiveness in our manufacturing sector.

UNEMPLOYMENT AND PRODUCTIVITY GROWTH

Popular discussions of productivity growth often warn that rapid increases in labor productivity are not as beneficial as they are cracked up to be. We are told that each gain in productivity reduces the demand for labor because it means that fewer work-hours are needed to produce a given output. As a result, according to this view, productivity growth allegedly creates unemployment. Second, and perhaps somewhat inconsistently, it is argued that if an economy's productivity growth lags behind that of other countries, it will lose jobs to foreign workers, its industry will suffer, and its exports will fall. However, the data do not support either of these conclusions, at least for the long run.

If the specter of growing long-run unemployment were a reality, we would expect that the enormous increases in output per labor hour in the United States, Great Britain, and Germany since 1870 would have had devastating effects on the demand for labor in these countries. After all, with productivity rising almost 13-fold in the United States during the last century, output per capita could have been kept about constant if employment were cut to one-thirteenth of its initial share of the U.S. population. Even with a 50 percent fall in the number of hours an average person works per year, we might expect unemployment of perhaps five-sixths of the U.S. labor force by now.

In fact, nothing of the sort has happened. Unemployment rates for Great Britain, the United States, and Germany going back to the 1870s indicate that

over this long period there was no upsurge in unemployment. Before 1914, unemployment in the three countries averaged about 4 percent of the labor force, while in the 1952 to 1973 period it averaged a bit more than 3 percent. Even though there has been a rise in unemployment throughout the industrial world recently, much of it is attributable to short-term influences and reorientation of public policy away from government intervention to reduce unemployment.

How have we maintained employment in the face of rising productivity? The answer, of course, is that output per capita has not remained constant. The demand for consumer goods and services, schools, hospitals, and factories has expanded explosively as productivity growth increased the purchasing power in the hands of the American public. That has sufficed to prevent any long-term increase in unemployment.

THE REAL COSTS OF LAGGING PRODUCTIVITY: LAGGING WAGES AND LIVING STANDARDS

As we have just seen, lagging productivity does not lead to a long-term rise in unemployment. For example, Great Britain, the most noted laggard in productivity growth among industrialized countries, has not suffered from unemployment problems markedly more serious than other countries'. In addition, its exports of goods and services have increased and its share of employment in manufacturing has pretty well kept up with other countries. These observations may make one wonder how Britain was able to score these apparent successes despite its comparatively poor productivity performance.

The secret, which also shows the true price Britain has paid, is found in its lagging real wages. In the 19th century, British workers were the best paid in Europe. According to one estimate, which admittedly is not very reliable, around 1860 an English worker's wages permitted the purchase of about 2 and 1/2 times as many goods and services as a German worker's. Yet by the end of the 1980s, the purchasing power of a German worker's wages was almost twice as great as that of a British worker's wages. In other words, in a little more than a century the relative position of workers in the two countries had almost been reversed.

How do lagging British wages relate to its productivity lag? The answer is straightforward. If Britain cannot compete on world markets by virtue of growing efficiency (productivity), it still can sell its products by providing cheap British labor. Of course, Britain does not volunteer to adopt low real wages; rather, market forces make it happen automatically, because inefficiently produced goods cannot be sold on the international marketplace unless those goods are produced by cheaper inputs. Hence, the invisible hand forces British wages to lag. Labor simply cannot extract higher wages from an economy that has little to offer.

A country with lagging productivity is likely to be condemned to become an exporter of cheap labor. That is the only way it can keep its industry viable, maintain its exports, and preserve domestic jobs. This is the real danger that the United States faces if its productivity performance is unsatisfactory for any substantial period of time.

The costs of failure in this arena are very high. Above all, lagging productivity growth must slow or bring to an end the rising living standards and rising real wages which have so long been a prime accomplishment of the U.S. economy. If there is no rise in output per worker, then it will be impossible to keep increasing the quantity of goods and services provided to each consumer. Indeed,

something of the sort has already happened. Since about the beginning of the 1970s, the growth in the purchasing power of an average American worker's hourly wage has fallen far behind its previous pace.

While U.S. productivity growth continues, as we have seen since the 1970s it has been below its historical rate.

REQUIREMENTS FOR INCREASED GROWTH

What can be done to increase the growth rate of production in an economy? Unfortunately, no one has a handy list of sure-fire recipes.

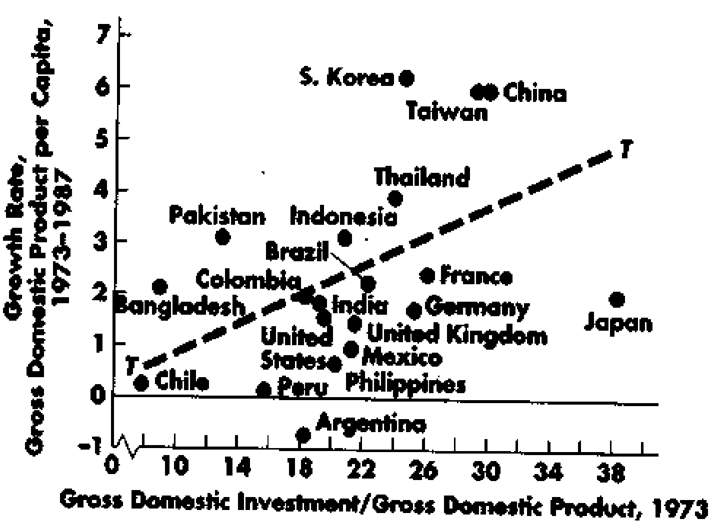
Growth can be attributed to a number of factors that no one knows how to explain: (1) *inventiveness,* which produces the new technology and other innovations that have contributed so much to economic expansion; (2) *entrepreneurship,* the leadership that recognizes no obstacles and undertakes the daring industrial ventures needed to move the economy ahead; and (3) *the work ethic* that leads a work force to high levels of productivity. No one really knows what features of economic organization and social psychology actually lead a community to adopt these goals, as Great Britain is said to have done at the beginning of the 19th century, as the United States is reputed to have done in the first half of the 20th century, and as some east Asia countries are apparently doing today. We do know, however, that growth requires two things that people can influence directly: a large expenditure on *capital equipment* and the devotion of considerable effort to *research and development* from which innovations are derived.

The *composition* of aggregate demand is a major determinant of the rate of economic growth. If a larger fraction of total spending goes toward investment rather than toward consumption, government purchases, or net exports, the capital stock will grow faster and the aggregate supply schedule will shift more quickly to the right.

Figure 37-8 shows, for a set of 20 countries on 4 continents, how investment in one period (in this case, 1973) is related to subsequent growth (from 1973 to

FIGURE 37-8

GROWTH AND INVESTMENT, 20 COUNTRIES



The graph shows, for 20 countries, how investment in 1973 was related to subsequent growth in per-capita GDP (1973 to 1987). The positive slope of the "trend line," *TT,* means that higher investment rates are associated with more rapid growth rates.

SOURCE: Angus Maddison, *The World Economy in the 20th Century* (Paris: Organization for Economic Cooperation and Development, 1989).

1987) in per-capita output. The graph confirms that the share of an economy's output devoted to investment does not, *by itself*, determine future growth. But higher investment rates clearly *are* associated with more rapid subsequent growth.

GROWTH WITHOUT SACRIFICING CONSUMPTION: SOMETHING FOR NOTHING?

We learned in Chapter 3 that the decision to invest involves a trade-off. By choosing to devote resources to future needs rather than current consumption, a society can achieve economic growth. In other words, through saving, the public gives up some consumption, which is the price it must pay for the accumulation of plant, equipment, and infrastructure. But while a substantial proportion of growth is embodied in these increased quantities of plant, equipment, and infrastructure, a very large proportion of the economy's growth is *disembodied*. That is, it is attributable to better ideas—to improved methods of finding and using the same quantities of resources. Everyone knows that this has in fact occurred. From the invention of the steam engine to that of the modern computer, our economy has benefited from a stream of inventions—some sensational, some more routine—which together have increased enormously the productivity of the nation's resources.

Embodied growth has two serious costs that disembodied growth avoids. First, embodied growth necessarily speeds up the use of society's depletable resources: its iron ore, its petroleum supplies, and its stocks of other minerals and fuels. Second, the greater the quantity of resources used in the productive process, the greater the quantity of wastes that must ultimately result.

So far, we have enjoyed substantial success in our efforts to achieve growth in output without commensurate increases in our use of resources. One statistical analysis, for example, attributes only about half of the growth in the United States to increased use of physical inputs. The remainder must be ascribed to improvements in technology as well as to increased education and skill of the labor force.[11]

One final remark on disembodied growth is in order. Economists are fond of pointing out that there is no such thing as a free lunch. Except in rare instances, improvements in technology are not "manna from heaven." They result, instead, from the work of scientists and technicians in government and industrial laboratories, from the labor of inventors in their basements or garages, and from the effort of management specialists studying the organization of factories and assembly lines. This means that labor (along with other resources) is diverted from other activities into the production of knowledge. *In a fully employed economy, the opportunity costs of investing in the discovery of new knowledge are the forgone consumption and physical investment of goods that would otherwise have been produced.* So even here, we cannot quite get something for nothing.

ON GROWTH IN POPULATION: IS LESS REALLY MORE?

We have just discussed some of the ways in which growth can be stimulated. There are also influences that impede growth. Many observers believe that

[11] Edward F. Denison, *Accounting for United States Economic Growth 1929–1969* (Washington, D.C.: The Brookings Institution, 1974).

expansion in population is one of the main obstacles to growth in per-capita income. Even though an expanding population means a rising labor force, there are reasons why output cannot keep up. Indeed, as we will see, there are reasons to believe that continuation of the rate of population growth of the past century is literally impossible.

The problem was first brought to the public's attention in 1798, when the Reverend Thomas R. Malthus (who was to become England's first professor of political economy) published *An Essay on the Principle of Population*. This book, which had a profound effect on attitudes toward population, argued that sexual drives and other influences induce people to reproduce themselves as rapidly as their means permit. Unfortunately, he said, when the number of humans increases, the production of food and other consumption goods generally cannot keep up.

The problem is the noted *law of diminishing returns* to additional labor using a fixed supply of land. This states that if we use more and more labor to cultivate a fixed stock of land, we will eventually reach a point at which each additional laborer will contribute less additional output than the previous laborer. Ultimately, as the labor force increases, output per worker will decline.

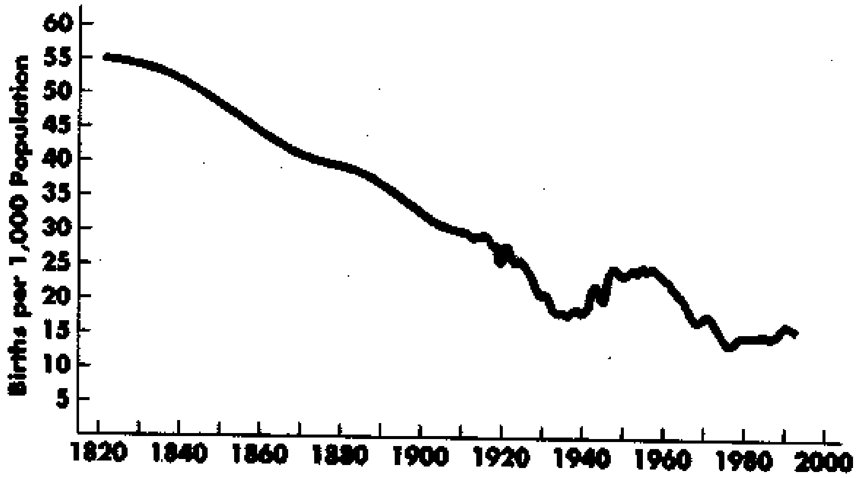
Malthus and his followers concluded that the tendency of humankind to reproduce itself must constantly exert pressure on the economy to keep living standards from rising. Wages will gravitate toward some minimal subsistence level—the lowest income on which people are willing to marry and raise a family. If wages are above subsistence, the population can and will grow. Thus, a wage that is above subsistence will set forces into motion that will drive wages down toward subsistence because of diminishing returns.

Sometimes, according to Malthus, the population will grow beyond the capability of the economy to support it. Then the number of people will be brought back into line by means that are far more unpleasant than a decrease in wages—by starvation and disease or by wars that produce the required number of casualties.

Later, in the 19th century and the first half of the 20th century, the gloomy Malthusian vision seemed to lose credibility. New technology and improved agricultural practices generally enabled the output of food and other agricultural products to increase faster than the population (at least in the wealthier, industrialized nations). In addition, it turned out that as living standards rose, people became less anxious to reproduce, and so the expansion of population slowed substantially. Figure 37-9 illustrates this trend in the United States for more than a century and a half. All in all, it began to look as though population growth constituted no significant threat—it was something with which human technological skills and ingenuity could cope.

More recently, however, there has been renewed concern over population. With improvements in medicine—notably, improved hygiene in hospitals, the use of such public-health measures as swamp drainage, insecticide spraying, inexpensive vaccination programs, and the discovery of antibiotics—death rates have plunged in the developing countries, especially for infants. At the same time, birth-control programs in most of these countries have, at least until quite recently, not been very successful. As a result, the populations of developing countries have continued to expand dramatically, eating up a good proportion of any output increases obtained through their governments' economic development programs. It has been widely concluded that significant improvement in living standards in the developing areas is impossible without substantial reductions in their population growth. But the "neo-Malthusians" go further than this, arguing that a rapid approach to birthrates so low that populations cease

FIGURE 37-9 AMERICAN BIRTHRATES, 1820–1993



This chart shows that birthrates in the United States have generally been declining since 1820.

SOURCE: U.S. Bureau of the Census, *Historical Statistics of the United States, Colonial Times to 1970, Part I* (Washington, D.C.: U.S. Government Printing Office, 1975); and U.S. Bureau of the Census, *Statistical Abstract of the United States* (Washington, D.C.: U.S. Government Printing Office, various editions).

expanding—that is, to *zero population growth*—is virtually a matter of life and death even for the most prosperous nations. It is illuminating to consider the logic of their argument.

THE CROWDED PLANET: EXPONENTIAL POPULATION GROWTH

Malthus used an argument that has caught many imaginations ever since:

> Population, when unchecked, increases in a geometrical ratio. Subsistence increases only in an arithmetical ratio. A slight acquaintance with numbers will shew the immensity of the first power in comparison of the second.[12]

EXPONENTIAL GROWTH is growth at a constant percentage rate.

In modern discussions, such a "geometric" growth pattern is referred to as **exponential growth,** "compounded growth," or "snowballing." Exponential growth is growth at a constant *percentage rate.* For example, at a 10 percent growth rate, a population of 100 persons will increase by 10 persons a year; but a population of a million persons will increase by 100,000 persons a year. The bigger the population, the more it will add annually. And each year's growth implies still faster growth in the following year. It is like a snowball rolling downhill, accumulating more snow the bigger it gets, thus expanding faster and faster all the time.

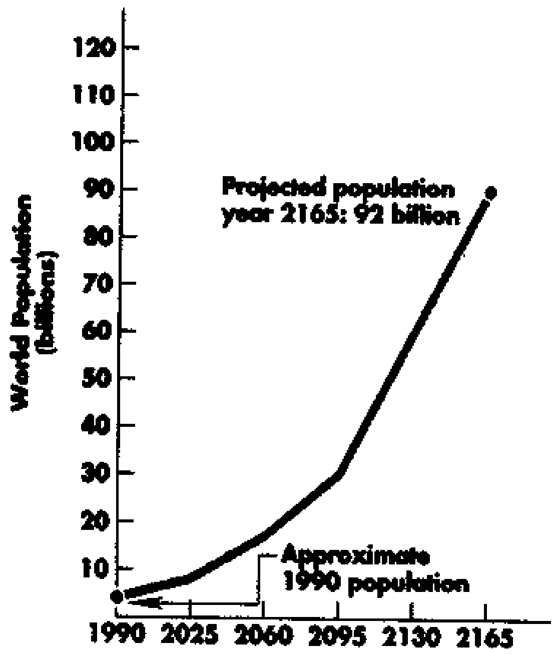
If the population doubles (grows 100 percent) in 35 years, it will quadruple (grow another 100 percent) in 70 years, increase 8-fold in 105 years, 16-fold in 140 years, and so on indefinitely. The doubling sequence 2, 4, 8, 16, 32, 64, and so on, is the basic pattern of exponential growth. Figure 37-10 shows how astronomical such a growth sequence is. Projecting the world's population 175 years into the future on the assumption that population will grow exponentially at about its current rate, the graph shows that by the year 2165 the population will exceed 90 billion—nearly 20 times as many people as there are today.

It turns out that in describing the consequences of exponential growth, Malthus was conservative. He did not begin to spell out the wonders and the

[12] Thomas R. Malthus, *An Essay on the Principle of Population* (New York: W. W. Norton, 1976).

FIGURE 37-10 PROJECTED GROWTH OF THE WORLD'S POPULATION IN 175 YEARS AT CURRENT RATE OF GROWTH

This figure shows the sensational acceleration of population growth *if* population expands exponentially.



horrors that his premise implied. Consider some calculations by one leading authority on population (who derived his conclusions simply by carrying through the arithmetic of exponential growth rates):

- If population were to grow at today's rates for another 600-700 years, every square foot of the surface of the earth would contain a human being;
- If it were to expand at the same rate for 1,200 years, the combined weight of the human population would exceed that of the earth itself;
- If that growth rate were to go on for 6,000 years (a very short period of time in terms of biological history), the globe would constitute a sphere whose diameter was growing with the speed of light.[13]

And none of this is conjecture. It is *sure* to come about *if* the present (exponential) rate of growth of the earth's population continues unabated.

Of course, none of this can really happen. Our finite earth just does not have room for that sort of expansion. If the number of humans continues to swell until it presses upon the earth's capacity, the process will ultimately be brought to a halt in a Malthusian apocalypse. Disease, famine, and war must finally put a stop to the expansion process. But there is a better alternative. People can choose to stop raising large families. And it can be done. Even in the developing nations, as we will see later in this chapter, the birthrate has recently been declining.

IS MORE GROWTH REALLY BETTER?

Many have questioned the desirability of faster economic growth as an end in itself, at least in the wealthier industrialized countries. Yet faster growth does

[13]Ansley J. Coale, "Man and His Environment," *Science*, vol. 179 (October 9, 1970), pp. 132–136 Copyright 1970 by the American Association for the Advancement of Science.

mean more wealth, and to most people the desirability of wealth is beyond question. "I've been rich and I've been poor—and I can tell you, rich is better," a noted stage personality is said to have told an interviewer. Most people seem to have the same attitude about the economy as a whole. To those who hold this belief, a healthy economy is one that is capable of turning out vast quantities of shoes, food, cars, and TV sets. An economy whose capacity to provide all these things is not expanding is said to have succumbed to the disease of *stagnation.*

Yet the desirability of further economic growth for a society that is already wealthy has been questioned on reasonable grounds. It is pointed out that the sheer increase in quantity of products has imposed an enormous cost on society in the form of pollution, crowding, and proliferation of wastes that need disposal; some economists claim it has had unfortunate psychological and social effects. It is said that industry has transformed the satisfying and creative tasks of the artisan into the mechanical and dehumanizing routine of the assembly line. It has dotted our roadsides with junkyards, filled our air with pollution, and poisoned our food with dangerous chemicals. The question is whether the outpouring of frozen foods, talking dolls, CD players, and headache remedies are worth the high costs to society.

Despite the costs of growth in terms of human and environmental damage, there is strong evidence that if the economy's total output were kept at its present level, the community would pay a high price over and above the loss of additional goods and services.

First, it is not easy to carry out a decision to prevent further economic growth. Mandatory controls are abhorrent to most Americans. We cannot *order* people to stop inventing means to expand productivity. Nor does it make any sense to order every firm and industry to freeze its output level, since changing tastes and needs require some industries to expand their outputs at the same time that others are contracting. But who is to decide which should grow and which should contract, and how shall such decisions be made?

Second, without continued growth it will be no easy matter to finance effective programs to improve the quality of life—programs such as environmental protection. To improve the purity of our air and water and to clean up urban neighborhoods, tens of billions of dollars must be made available every year. Continued growth can permit the required resources to be provided without any reduction in consumer goods. But without such growth, we may actually be forced to cut back on our programs to protect the environment. Society could thus end up with fewer goods and a worse environment.

Finally, zero economic growth may seriously hamper efforts to eliminate poverty both within our economy and throughout the world. Much of the earth's population today lives in a state of extreme want. And though wealthier nations have been reluctant to provide more than token amounts of help to the less developed countries, less wealth means there would be even less to share. So the best hope for improved living standards in the impoverished countries of Africa, Asia, and Latin America lies in continued increases in output.

PART 2: PROBLEMS OF THE LESS DEVELOPED COUNTRIES

A substantial proportion of the world's population lives in areas whose average per-capita GDP is $4,000 or less per year, evaluated (as well as is possible) in current U.S. prices. And Table 37-1 shows that there are countries in which annual per-capita GDP is less than $1,000. Even after adjustment for differences

TABLE 37-1

1994 PER-CAPITA GDP IN DEVELOPED AND LESS DEVELOPED COUNTRIES[a]

	Per-Capita GDP
Developed Countries	
United States	25,850
United Kingdom	17,980
Sweden	18,580
Less Developed Countries	
Bolivia	2,370
China	2,500
Egypt	2,490
Ethiopia	380
Haiti[b]	870
India	1,360
Tanzania	750

[a]Measured in U.S. dollars, using PPP exchange rates. See text footnote 6 for an explanation.

SOURCE: Central Intelligence Agency, *The World Factbook 1995* (Washington, D.C.: CIA, 1995).

in measurement of GDP in the United States and these extremely poor countries, their annual per-capita GDP figures are still under $1,500.

To us, residents of an economy that offers an average GDP more than 15 times as high as this, such a figure is not only likely to seem incredible, it is all but incomprehensible. Yet these figures do *not* represent the living standards of a small group of outcasts from their own societies. Rather, they are *typical* of many people living in Asia, Africa, and Latin America.

What is life like in such circumstances? No brief description can really bridge the gulf between us and them, but it can provide a glimpse of a way of life few of us want to share.

Inhabitants of many of the LDCs live with their large families in one-room shanties or apartments; their water supplies are scanty, polluted, and often miles from home; their only source of energy is that of human and beast; in good years sparse harvests are wrung from miserable soil, with starvation threatened perhaps every five years when the rains do not come and the crops fail. Low production permits no food reserves, and the old, the infirm, and the very young are likely to perish in famine years.

The life of a man in an LDC is hard enough, with its poor nutrition, lack of work equipment, and frequency of debilitating diseases. But his life is luxurious compared with that of a woman in these countries. She is usually married by the age of 14, bears 8 or 10 children with minimal medical assistance, and by 35 is often beset by chronic disease and aged far beyond her years. If (as is true of some 80 percent of the population) she inhabits a rural area, she may have to walk miles every day to fetch water for the family. She sews the family's clothes and cooks its meals. There is not enough money for pre-ground flour, so part of

TABLE 37-2 INFANT MORTALITY AND LIFE EXPECTANCY IN DEVELOPED AND LESS DEVELOPED COUNTRIES, 1995

	Infant Mortality (deaths per 1,000 live births)	Life Expectancy at Birth (years)	
		Males	Females
Developed Countries			
United States	8.0	72	79
Germany	5.8	73	79
Sweden	4.8	76	81
Less Developed Countries			
Bolivia	71.0	59	62
China	44.0	67	70
Egypt	62.0	62	65
Ethiopia	120.0	48	52
Haiti	74.0	55	58
India	74.0	60	60
Tanzania	92.0	47	50

SOURCE: Population Reference Bureau, Inc., *World Population Data Sheet, 1995.*

a woman's daily work is to pound the grain by hand for food for the family—perhaps an additional two hours of hard labor. She also tends the gardens that produce food for the family, although, except in Moslem countries where women are sequestered, she is also expected to put in a full day in the fields during the six months of the agricultural season.

Another duty of the woman in an LDC is to bring produce, wood, or whatever she has to trade to market a couple of times a week, and she must often walk as many as 10 miles each way with bundles as heavy as she can carry on her back or on her head. She has no respite in the raising of her children, since they are likely not to have a school to attend when they are well or a hospital to go to when they are sick. It is no wonder that she ages so much faster than a woman in our society.

Table 37-2 gives the percentage of infant deaths for each 1,000 live births and the average life expectancy of a newborn child in a sample of countries ranging from the most underdeveloped to the most affluent. The contrasts are dramatic. In Ethiopia, 120 babies die of every 1,000 that are born, while the comparable figure in Sweden is under 5. In many countries people survive only until their 40s or 50s, while in Sweden the average life expectancy is 76 years for men and 81 years for women. There is little question about the quality of life in less developed lands.

RECENT TRENDS

Despite population increases, some LDCs, particularly in the Far East, have succeeded in breaking out of the stagnation trap. In those successful economies, if

growth were to continue as it was in the 1970s, an average family could look forward to a doubling of its living standards in less than 30 years. Or put another way, in earlier decades standards of living were increasing faster than they did in the United States in the 19th century!

While in the 1970s such good news applied to a number of LDCs in several parts of the world, the 1980s were not so favorable. Many of the LDCs had drastic debt problems. In addition, their per-capita incomes tended to fall behind those of the developed countries. Largely because their population growth has been faster, the percentage growth rates in per-capita incomes have been lower in the LDCs. But even if the *percentage* increases in their per-capita incomes had been similar, *absolute* incomes would have continued to rise more quickly in the richer lands. For example, where per-capita income is $100 a year, a 2.5 percent growth rate translates into a $2.50 annual improvement. However, where per-capita income is $5,000 a year, the same 2.5 percent rate of growth adds $125 a year to the income of the average person.

A few numbers will indicate how discouraging the relative performance of the LDCs as a group has been. A study of some 70 countries by a group of noted economists showed how much the disparity between the LDCs and the other two groups has been increasing.[14] The average annual growth rates of real GDP were 3.1 percent (compounded) for the industrial countries, 3.0 percent for the middle-income countries, and only 1.5 percent for the LDCs. With average growth rates half as big as those for the more affluent economies, the LDCs as a group fell further behind the rest of the world for much of the period since World War II.

IMPEDIMENTS TO DEVELOPMENT IN THE LDCs

No one has produced a definitive list of causes of the poverty of the LDCs, just as no one can pretend to have produced a foolproof prescription for its cure. Yet there is general agreement on the main conditions contributing to the economic problems of LDCs. These include a dearth of physical capital, rapid growth of populations, lack of education, unemployment, and social and political impediments to business activity. Let us consider each of these.

SCARCITY OF PHYSICAL CAPITAL

The LDCs are obviously handicapped by their lack of modern factories and machinery. In addition, they lack infrastructure—good roads, railroads, port facilities, and so on. But capital is not easy to acquire. If it is to be provided by the populations of the LDCs themselves, they must save the required resources—that is, as we noted earlier, they must give up consumption in order to free the resources needed to build plants, equipment, and roads. That is comparatively easy in a rich community, where substantial saving still leaves the public well off in terms of current consumption. But in an LDC, where malnutrition is a constant threat, the bulk of the inhabitants cannot save except at enormous sacrifice

[14] Robert Summers, Irving B. Kravis, and Alan Heston, "Changes in World Income Distribution," *Journal of Policy Modeling*, vol. 6 (May 1986), pp. 237–269.

to their families. Moreover, in many of the LDCs, tradition imputes little virtue to investment in business, so that even the wealthy are not terribly anxious to put their savings into productive equipment. Thus:

Because poverty makes saving difficult and traditions often do not encourage investment, the LDCs' growth rates of domestically financed capital are lower than those in the developed countries.

One way to help matters is to obtain the funds for investment from abroad. In recent decades a considerable share of the resources going to the LDCs has come from foreign aid programs. While some such resources have been squandered, informed observers generally agree that the waste has not been spectacularly great. They conclude that these capital transfers from the rich countries to the poor are a step in the right direction.

Capital can also be transferred to an LDC when a private firm invests money to build a factory or to explore for oil. This increases the firm's profits and benefits the LDC. In earlier days, it sometimes gave an unacceptable degree of political influence to the foreign firms, particularly when the LDC was a colony of an industrial country. In recent years this difficulty may have become more rare. Nowadays, it is more often the outside firm that is afraid of the government of the LDC rather than vice versa. Foreign proprietors frequently fear rigid governmental controls. Another fear is outright expropriation—where the LDC's government simply takes over the foreign firm's property, with or even without compensation, because of the hostile attitudes that local residents may hold toward large foreign companies.

POPULATION GROWTH

Population growth is often described as the primary villain in the LDCs. We have already noted that their populations grow far more rapidly than those in the wealthier countries. And although growth rates have recently been declining in many of the less developed countries, overall, the population of the LDCs is expanding at a rate that will double in less than 30 years, requiring a doubling of housing, schools, hospitals, and so on—a heavy real cost for an LDC.

Table 37-3 tells the story. For the LDCs shown, the average annual growth rate of population is 10 times the average rate of the industrialized countries. Clearly, the more closely the growth in population approximates the growth in national income, the more slowly standards of living will rise, since there will be that many more persons among whom the additional product must be divided.

As we noted earlier, the growth in population has been stimulated by improvements in medical care, which have reduced death rates spectacularly. Today, in some areas, death rates (ratio of deaths to population) are only one-quarter or one-fifth as high as birthrates. In the past it was not unusual for half of a nation's children to die before the age of 20. In many countries today this is true of only some 4 percent of the population.

Not all LDCs suffer from serious population problems such as those that plague India, Indonesia, and Egypt. In fact, many countries in Africa and Latin America have populations so small that they are denied economies of larger-scale communication and transportation. The economy of a sparsely settled country whose electric power and telecommunication lines must traverse great unpopulated areas is under a costly handicap.

Governments in a number of LDCs have been struggling to find workable ways to cut population growth. Programs set up to distribute contraceptives and propaganda against large families have achieved modest success; but in some

TABLE 37-3

NATURAL INCREASE IN POPULATION
(BIRTHRATE MINUS DEATH RATE)
IN DEVELOPED AND LESS DEVELOPED COUNTRIES, 1995

	Births Minus Deaths (as percentage of population)
Developed Countries	
United States	0.7
Germany	−0.1
Sweden	0.1
Less Developed Countries	
Bolivia	2.6
China	1.1
Egypt	2.3
Ethiopia	3.1
Haiti	2.3
India	1.9
Tanzania	3.0

SOURCE: Population Reference Bureau, Inc., *World Population Data Sheet, 1995.*

countries with particularly severe population problems the governments have been dissatisfied with the results of these voluntary efforts and have opted instead for compulsory sterilization. This has aroused the anger of the public and in some cases has led to the downfall of the governments.

EDUCATIONAL AND TECHNICAL TRAINING

Everyone knows that educational levels in the LDCs are much lower than those in wealthier countries. There are fewer graduates of elementary schools, far fewer graduates of high schools, and enormously fewer college graduates. The percentage of the population that is literate is much lower than that in industrialized nations. The issue is how much of a handicap this constitutes for economic growth.

The evidence is that general education, as distinguished from technical (trade) schooling, makes considerably less difference for economic growth than is often believed. For example, the number of jobs that clearly require secondary (high school) education rarely exceeds 10 percent of the labor force. Various studies that have investigated whether there is a statistical relationship between the economic growth of an economy and its typical educational level have found only weak correlations between the two.

This is not meant to imply that education is worthless. On the contrary, it obviously offers many inherent benefits, which need not be discussed here. But it does suggest that if a government invests in education *purely as a means to stimulate economic growth,* only a very limited outlay on *general* education beyond the achievement of literacy may be justifiable.

Matters are quite different when we turn to technical training. There is apparently a high payoff to the training of electricians, machinists, draftsmen,

construction workers, and the like. While the number of persons involved need not be very high in proportion to the population, the role played by such specialists is crucial. One of the main inhibitions to adequate training in these areas is that in many countries such skills are held in low esteem and considered inferior to training in the liberal arts. Consequently, technical education is often handicapped by low budgets, low teacher salaries—which discourage good people from entering the field—and the prejudice of potential students against such fields.

Training in improved farming methods also has a great deal to contribute. As one leading observer, the late Nobel Prize-winner Sir W. Arthur Lewis, remarked, "... no impact can be made on mass [LDC] living standards without revolutionizing agricultural performance."[15] There seem to be no easy ways to provide the necessary education to farmers who cannot spare the time to attend schools. Religious beliefs often lead parents to object to schooling of their children, particularly of girls; truly literate and knowledgeable teachers are often almost impossible to find in any substantial numbers; and children who do complete schooling often leave the farms for the cities.

UNEMPLOYMENT

Growth in the LDCs has been accompanied by increased unemployment as population has shifted from agriculture to the cities. Increased schooling has stimulated this migration, as has unionization, which has often produced a huge gap between urban and rural wages. Government investment policies have also favored construction of schools, hospitals, and other facilities in the cities, and as a result, large numbers of migrants have entered the cities to swell the ranks of the unemployed. The unemployment rate among young, urban workers has been particularly high; indeed, rates as high as 50 percent are not unheard of.

These figures are compounded by the phenomenon of *disguised unemployment*. For example, ten persons may do a job for which only six are needed. The statistics would show no unemployment among the ten workers, even though four of them really contribute nothing to output. Some observers believe that this is such a widespread problem in rural areas that even a substantial reverse migration of the urban unemployed back to the farms would add very little to production, at least in some of the LDCs.

An important consequence of all this is that in many LDCs unemployment may not be accompanied by any substantial reduction in output, in contrast to the situation in industrialized economies. But unemployment in the LDCs is a serious social problem.

SOCIAL IMPEDIMENTS TO ENTREPRENEURSHIP

As we noted earlier in this chapter, one of the magic ingredients of economic growth is *entrepreneurship*. It is clear that the LDCs need entrepreneurs if their economies are to grow rapidly. But in many of these economies there are serious inhibitions to entrepreneurship. Traditional social values often accord relatively low status to business activity. Indeed, those traditional values even prevent businesses from seeking ways to attract and please their customers and their

---

[15] W. Arthur Lewis, *Development Economics: An Outline* (Morristown, N.J.: General Learning Press, 1974), p. 25.

work forces. In addition, high positions in business in many LDCs are often determined by family connections and inheritance, not by ability.

In the LDCs, growth will be inhibited until customs can be modified to increase the social status of economic activity, to make it respectable for private business people and managers of public enterprises to do their best to attract business and increase productivity, and to assign responsibility on the basis of ability rather than family connections.

GOVERNMENT INHIBITION OF BUSINESS ACTIVITY

In addition to social impediments to business, the political situation in the LDCs often is detrimental to business success. Business is not helped by unstable governments or by the uncertainty that accompanies such an environment, especially if there is a high likelihood of revolution or internal warfare. Foreign investment will be discouraged where there is fear of expropriation or of unstable currencies that may fall in value and wipe out hard-earned profits. And indigenous business people may live in fear of nationalization or even imprisonment—possibilities that are not likely to encourage investment.

In the normal course of events, governments in the LDCs are often inclined to interfere with business activity in a variety of ways that seem relatively innocuous—but whose effects can be deadly. Price controls are often imposed at levels that make the controlled activity totally unprofitable and cause it to wither. Licenses and other direct controls are frequently administered by incompetent bureaucrats who tie up business activity in red tape. As a matter of prestige of the currency, exchange rates are often set so high that exports from the LDC cannot compete on the world market. The governments sometimes expropriate and seek to operate foreign firms before they have trained native personnel to run them. In short:

Poorly conceived economic policies can impede business activity and hence economic growth in the LDCs. But, then, it must be admitted that the LDCs have no monopoly on foolish economic policies!

HELP FROM INDUSTRIALIZED ECONOMIES

We have just seen that the two primary needs of the LDCs are technical skills and capital resources. Happily, these are precisely the things that the more prosperous nations are in a position to offer. We have the trained teachers, classrooms, laboratories, and equipment necessary to provide an education of the highest quality to students from the LDCs. However, there is a danger here that has received a great deal of attention, the so-called *brain drain*. This refers to the temptation for students from LDCs to try to stay in the countries where they have studied and enjoy the higher living standards, rather than to return home where their abilities are needed so badly.

There are several ways to deal with this. For example, one can require students to return to their homelands for at least some number of years after completion of the educational program or offer higher wages as an incentive for trained persons from the LDCs to return. The large number of doctors, teachers, and other skilled personnel from LDCs who are seeking jobs in the developed countries suggests that this is not a negligible problem.

A second major contribution that the wealthier countries can make to the LDCs is to supply them with trained technicians and technical advice from their

own populations. Such counseling and personnel can be very helpful as a temporary measure, but in the long run can prove detrimental if provision for the training of local personnel for the ultimate replacement of the foreign technicians and advisers is not built into the program.

Third, the world can help the LDCs through research. One of the hardest problems in heavily populated rural areas of Asia and Africa is what to do about persistent drought. New dry-farming techniques are badly needed and until some are discovered, poverty will increase as the population grows. An international research organization devoted to food production in problem areas in the LDCs would have much to contribute.

Fourth, the developed countries can help by encouraging freedom of trade and investment and the elimination of trade barriers. Exports of sugar, meat, cotton, and other agricultural products are inhibited by industrialized countries' tariffs and other restrictions. LDCs would also benefit substantially from a lifting of tariffs and quotas upon imports into industrial countries of processed or manufactured goods. Such restrictive measures make it difficult for LDCs to export anything but raw materials and impede their industrialization and modernization. All in all, increased freedom of trade is a top priority for the LDCs. It will probably also prove beneficial to the wealthier countries, as improved incomes in the LDCs make the latter better customers for the exports of the former.[16]

A last, and very important, type of assistance from the developed to the less developed countries takes the form of money or physical resources provided either as loans made on favorable terms or as outright grants (gifts).

LOANS AND GRANTS BY THE UNITED STATES AND OTHERS

Since World War II a number of countries have provided capital to the LDCs. The International Bank for Reconstruction and Development (which, along with a closely related institution, the International Development Association, makes up the *World Bank*) was created largely for this purpose. It has 177 member countries, each providing an amount of capital related to its wealth; for instance, the United States has contributed about 20 percent of the total, while Japan and Germany have supplied 14 percent and 11 percent, respectively. The bank makes loans that finance its bonds and has acted as guarantor of repayment to encourage some private lending. Since its inception the bank has approved loans totaling $333.8 billion to 162 countries, mostly for infrastructure, dams, communications, and transportation. In addition, it has provided technical assistance and planning advice.[17]

United States loans and grants have exceeded the total given by all other countries and international agencies, with U.S. interest and repayment terms generally far more generous than those of other governments. However, the bulk

---

[16]Not everyone agrees with this conclusion. There are those who have argued that it is bad for LDCs to participate in international trade because it weakens their capacity to develop as self-reliant, mature economies. It is believed that new manufacturing industries in the LDCs will not take off without protection from foreign competition, that development of raw material exports creates a politically powerful vested interest that inhibits manufacturing, and that foreign participation in trade and production of exports inhibits domestic investment and the development of local entrepreneurship.

In this view, LDCs are therefore held back by international trade and they would do better to integrate regionally and develop their own home markets without foreigners, who also bring unsuitable habits, tastes, and technology and impart crippling inferiority complexes to the LDCs.

[17]World Bank, *The World Bank Annual Report 1994* (Washington, D.C., 1994).

of the assistance provided by the United States has gone to a small number of countries, such as India, Pakistan, South Korea, and Turkey.

During the 1960s, our expenditures on aid ran to more than $3 billion per year. In the past decade, expenditures on foreign aid have become less popular politically and have gone down from about 0.5 percent of U.S. GDP in 1965 to about 0.3 percent of GDP in 1992.[18] France, Great Britain, Germany, Japan, and other industrialized countries now provide approximately $70 billion per year, which is about 1.0 percent of their combined GDPs.

Many economists have advocated greater generosity in our assistance to LDCs. It is argued that, by making those countries more stable economically and politically, we can contribute to our own economic tranquility. By increasing the LDCs' power to buy and sell, we contribute to the prosperity of the entire world.

CAN LDCs BREAK AWAY FROM POVERTY?

It is easy to jump to the conclusion that the economic problems of the LDCs are staggering and that the prospects of their ever catching up with the industrialized countries are negligible. Yet a number of LDCs and former LDCs have made enormous progress. In Africa, Nigeria, Senegal, and Ghana increased their GDPs during the late 1980s at a rate of about 5 to 6 percent a year, which is considerably faster than their population growth. In the Americas, Costa Rica's performance has been comparable. Even more striking is the expansion of output in a number of places in the Far East—particularly Hong Kong, Taiwan, South Korea, and Singapore, where prosperity is unprecedented and economic activity is expanding at an astonishing rate. Here, per-capita GDPs have been growing at a rate of 6 percent a year and more. These economies have progressed to a point that makes it inappropriate to continue to classify them as LDCs.

But the most impressive case is that of Japan. Many of your professors will remember clearly when U.S. business feared the flood of goods produced by cheap Japanese labor, and when the label "made in Japan" suggested inexpensive and shoddy merchandise. From one of the world's impoverished countries, Japan has risen to one of the world's richest. Its goods are now feared by American manufacturers not because they are produced and sold so cheaply, but because their quality is so high. Japanese cars and sophisticated electronic equipment find a ready market in the United States. And as a result, per-capita income in Japan has surpassed that in Great Britain.

Clearly, a less developed country need not lag behind forever.

[18]Bureau of the Census, *Statistical Abstract of the United States, 1995*, 115th ed. (Washington, D.C.: U.S. Government Printing Office, 1995), pp. 451 and 810.

SUMMARY

1. Productivity growth over the past century has made a tremendous contribution to *standards of living*. Real U.S. per-capita income is more than 9 times as large as it was in 1870. Never in previous history have economic conditions improved so much.
2. For the first time in history, famine is no longer a constant threat in the world's industrialized countries. That is because productivity in agriculture has increased greatly. In 1800 about 90 percent of the U.S. labor force was needed to feed the nation poorly. Today, less than 3 percent of the labor force works on farms and yet produces great abundance.
3. Because of compounding, over longer periods a small increase in rate of productivity growth can

make an enormous difference in the economic well-being of a nation. This is one of the **Ideas for Beyond the Final Exam.**

4. There is evidence suggesting that at least a small set of the world's leading economies are converging toward similar living standards and similar productivity levels.
5. All nations learn about new technological developments from one another. The international spread of inventions means that research in one country also benefits inhabitants in much of the rest of the world.
6. Since the 1970s, there has been a substantial slowdown in productivity growth in most industrial countries.
7. The growth of productivity in U.S. manufacturing shows no sign of a long-term slowdown.
8. The growth rate of U.S. productivity has for many decades been slower than that of a number of other industrial countries. However, at least in part, that reflects the fact that the United States is still the world's productivity leader; so that while we have much to learn from others, they have even more to learn from us.
9. The share of the U.S. labor force employed in the services has increased substantially. But so has that of every industrial, free-market economy. A major cause is probably the rapid rise of manufacturing productivity, which means that fewer workers are needed in the manufacturing sector.
10. Lagging productivity holds back real wages and per-capita incomes in a country. However, in the long run it will generally not cause unemployment or inability to export enough to pay for the nation's imports.
11. Increases in growth depend heavily on entrepreneurship, accumulation of capital equipment, and research and development.
12. Saving is necessary for the accumulation of resources with which to produce factories, machinery, and other capital equipment. Thus, saving is a critical requisite for growth, particularly in less developed countries.
13. A rapidly rising population poses a threat to growth of per-capita income.
14. On our finite planet, **exponential growth** (growth at a constant percentage rate) is, in general, impossible except for relatively brief periods of time.
15. Many observers argue that even if continued growth does not lead to catastrophically rapid depletion of resources (as some have predicted), it nevertheless has some questionable effects such as pollution and overcrowding.
16. Those who favor growth argue that without it there is no chance of ridding the world of poverty.
17. Standards of living in many *LDCs* are extremely low; per-capita incomes that are equivalent to $1,500 a year are not uncommon. Life expectancy is low and daily living is very difficult, particularly for women.
18. The gap between *per-capita* incomes in the less developed and the industrialized countries has continued to widen.
19. Growth in the LDCs is impeded by shortages of capital caused by poverty, traditions that do not encourage investment, rapid population growth, poor education, unemployment, lack of *entrepreneurship*, and government impediments to business.
20. Industrialized countries can help the LDCs by providing capital through loans and grants, by offering training and education to people from those lands, and by encouraging freedom of trade with the LDCs.
21. In the period after World War II many countries provided large amounts of money to the LDCs in the form of loans and grants.
22. Several international organizations, most notably the *World Bank*, have been organized to provide economic assistance to the LDCs.

KEY TERMS

Labor productivity
Standard of living
Exponential growth
GDP per capita
Deindustrialization
Embodied growth
Disembodied growth
Less developed countries (LDCs)
Disguised unemployment
Entrepreneurship
Brain drain
World Bank

QUESTIONS FOR REVIEW

1. Try to describe what family budgets were like 120 years ago when U.S. income per person (GDP per capita) was about one-ninth as high as it is today.
2. List some of the inventions that have increased agricultural output in the past century.
3. List some of the inventions that have increased manufacturing output in the last century.
4. List some of the new consumer products of the past century. Which of them became generally available only since World War II?
5. If growing productivity has vastly reduced the amount of labor needed to produce a given output, why has it not caused massively growing unemployment?
6. Which do you think are more similar?
   a. Production methods in a U.S. and a German factory today
   b. Production methods in a U.S. factory today and the methods used in that same U.S. factory 25 years ago
7. Which do you think has the higher total GDP, Pakistan or Luxembourg? Which has the higher per-capita GDP? In which do you think people are better off economically?
8. Suppose population grows at a constant, exponential rate and doubles every 12 years. How many times will it have grown in 36 years? How many years does it require to expand to 32 times its initial level?
9. Can you think of any innovations that permit growth without proportionate increases in use of inputs?
10. List as many undesirable consequences of growth as you can think of.
11. Are the undesirable consequences of growth more likely to be considered serious in a less developed country or in an industrialized country? Why?
12. To many families living in less developed countries, an income equivalent to $2,000 per year is considered a high standard of living. Can you make up a budget for a U.S. family of four earning $2,000 a year?
13. Explain how it is possible for the per-capita income of an LDC to grow at a faster rate than that in the United States and yet for the dollar difference between the incomes of average families in both countries to increase. Can you give a numerical example showing how this happens?
14. It has been noted that crime overlords who organize drug empires and the managements of law firms that specialize in stimulation of litigation are often, in fact, successful entrepreneurs. Discuss whether these persons contribute to the growth of their economies.

CHAPTER 38

COMPARATIVE ECONOMIC SYSTEMS: WHAT ARE THE CHOICES?

They pretend to pay us, and we pretend to work.
Polish folk definition of communism

There is no worthy alternative to the market mechanism as the method for coordinating economic activities.
Leonid Abalkin (Deputy Prime Minister of the Soviet Union, 1989)

There can be little doubt that, when the history of the late 20th century is written, the fall of communism will stand out as its defining event. For decades, the rivalry between Western capitalism and Soviet-style socialism had dominated the world's geopolitical scene. This competition had important economic, political, and military dimensions. But economic disarray in the planned socialist economies ultimately took its toll. When the end came, it came quickly. Within a few years the Iron Curtain fell, the Soviet Union disintegrated, and the economies of Eastern Europe began a painful *transition* from socialist planning to capitalist markets.

The transition process starkly illustrates that a society can in fact *choose* its economic system. In somewhat less dramatic fashion, the same thing is now going on in other parts of the world, including China, India, and parts of Latin America.

Here in the United States, we tend to take economic institutions as given and immutable. But they are not. In fact, there are many ways to practice capitalism, as the differing economic structures of Japan and Western Europe illustrate. A century ago, our country had no Social Security system, no income tax, no central bank, no antitrust laws, and hardly any labor unions. More than likely, the structure of the U.S. economy will change at least as much in the next century as it did in the last. But in which directions?

In this chapter we examine how a society might choose among *alternative economic systems*. The first parts of the chapter analyze the two major choices that every society must make: Should economic activity be organized through *markets* or by government *plan*? And should industry be *privately* or *publicly* owned? As we shall see, there are arguments on both sides of each question; so it is not surprising that different countries at different times have made different choices.

In the last sections of the chapter, we describe some of the actual choices that have been made in the contemporary world. We examine changes in Russia and in the People's Republic of China, two socialist nations that are transforming themselves in quite different ways. Then we turn to a notable example of a successful capitalist country that does things rather differently than we do: Japan. Does the United States have much to learn from the experiences of these other countries, or they from us? Read this chapter and then decide.

ECONOMIC SYSTEMS: TWO IMPORTANT DISTINCTIONS

Economic systems can be distinguished along many lines, but two are most fundamental.

The first is: *How is economic activity coordinated—by the market or by the plan?* The question does not, of course, demand an "either, or" answer. Rather the choice extends over a wide range from pure laissez-faire to rigid central planning. Society must decide which decisions it wants made in markets by individual businesses and consumers acting in their own self-interests, and which decisions it wants centrally planned so that businesses and consumers act more "in the national interest."

The second crucial distinction among economic systems concerns *who owns the means of production*. Specifically, are they privately owned by individuals or publicly owned by the state? Again, there is a wide range of choice and, to our knowledge, there are no examples of nations at either the **capitalist** extreme where all property is privately owned or at the **socialist** extreme where no private property whatever is permitted.

CAPITALISM is a method of economic organization in which private individuals own the means of production, either directly or indirectly through corporations.

SOCIALISM is a method of economic organization in which the state owns the means of production.

For example, while most industries are privately owned in the United States, a few are not. And many business owners face restrictions on what they can do with their capital. Automobile companies must comply with environmental and safety regulations. Private communication and transportation companies may have both their prices and conditions of service regulated by the government. And even in China, where large enterprises are publicly owned, anyone who can afford it can own a car, a bank account, or even a small business.

People tend to merge the two distinctions and think of capitalist economies as those with both a great deal of privately owned property *and* heavy reliance on free markets. By the same token, socialist economies typically are thought of as highly planned, as were those in the former Soviet bloc. However:

While there is an undeniable association between private ownership in a country and the degree to which it relies on markets, it is a mistake to regard these two features as equivalent. Socialism can exist with markets and capitalism can exist with rigid state planning. So, in thinking abstractly about a society's *choice* among economic systems, it is best to keep the two distinctions separate.

History holds examples of planned, capitalist economies—such as Germany under Hitler, Italy under Mussolini, and Argentina under Juan Peron. To a much lesser extent, Japan and the other "Asian tigers" have also planned their capitalist economies—apparently with great success.

We are not trying to suggest that capitalist economies are typically as heavily planned as socialist ones. In fact, they are not. But it is useful to view the choices

between planning and markets and between capitalism and socialism as two choices, not one. We take them up in turn.

THE MARKET OR THE PLAN? SOME ISSUES

The choice between *planning* and reliance on *free markets* requires an understanding of just what the market accomplishes and where its strengths and weaknesses lie. Since these issues have been discussed in earlier chapters, we can review them quickly and proceed to a comparison with central planning.

What goods to produce and how much of each. In a market economy, consumers determine which goods and services will be provided, and in what quantities, by registering their dollar votes. Items that are overproduced will fall in price, while items in short supply will rise in price. These price movements act as *signals* to profit-seeking firms, which then produce more of the goods whose prices rise and less of the goods whose prices fall. This mechanism is called **consumer sovereignty.**

CONSUMER SOVEREIGNTY means that consumer preferences determine what goods shall be produced and in what amounts.

Of course, consumers are not absolute monarchs, even in market economies. Governments interfere with the price mechanism in many ways—taxing some goods and services and subsidizing others. Such interferences certainly alter the bill of goods that the economy produces. We have also learned that in the presence of externalities the price system may send out false signals, leading to inappropriate levels of output for certain commodities.

Under central planning, the bill of goods that society produces is selected by planners rather than by consumers. Which system works better? Certainly, consumer sovereignty can lead to some bizarre products, the kinds of things that social reformers find offensive: designer jeans, junk food, low-quality television programming, and the like. But few of us would prefer the old Soviet-style system, which seemed incapable of providing what people wanted. Chronic shortages of ordinary consumer goods were a consistent feature of life in Eastern Europe and the Soviet Union—and one of the major causes of discontent.

How to produce each good. In a market economy, firms choose the production technique, guided by the price system. Inputs that are in short supply will be assigned high prices by the market. This will encourage producers to use them sparingly. Other inputs whose supply is more abundant will be priced lower, which will encourage firms to use them.

Once again, the same two qualifications apply: Government taxes and subsidies alter relative prices, and externalities may make the price system malfunction. But on the whole, the market system has no serious competitor as an engine of productive efficiency, as the formerly socialist economies learned.

Planned economies can allow either plant managers or central planners to choose the production techniques. In the former Soviet Union, plant managers had little discretion—which led to monumental inefficiencies including production curtailments, poor quality, and high costs. Indeed, such problems were among the most serious weaknesses of socialist planning. The truth is that no incentive system has yet been designed that can match the profit motive of competitive firms for keeping costs down.

How income is distributed. In a market economy, the price system determines the distribution of income among individuals by setting the levels of wages, interest rates, and profits. As we have stressed, there is no reason to expect the

resulting income distribution to be ethically appealing. And, in fact, the evidence shows that capitalist market economies produce sizable inequalities.

Income inequality is certainly one of capitalism's weak points. However, the government has many ways to mitigate inequality without destroying either free markets or private property (for example, through redistributive taxes and transfer payments). In practice, governments in all market economies intervene in the marketplace to alter the distribution of income.

So do governments in planned economies, only more so. For example, they may try to tamper directly with the income distribution by having planners, rather than the market, set relative wage rates. But doing so leads to shortages and surpluses of particular types of labor. So even in the former Soviet Union relative wages were established more or less by supply and demand, and there was considerable inequality in the distribution of wages.

Economic growth. The rate of economic growth depends fundamentally upon how much society decides to save and invest. In a free-market economy, these decisions are left to private firms and individuals, who determine how much of their current income they will consume today and how much they will invest for the future. Once again, however, government policies can influence these choices by, for example, making investment more or less attractive through tax incentives.

Planned economies have more direct control over their growth rates because the state can determine the volume of investment. They can therefore engineer periods of very high growth if they choose to—an option which both Stalin's Russia and Mao's China exercised successfully, though at great human cost. More recently, Singapore's planned market economy has grown rapidly by forcing individuals to save. One may wonder, however, whether rapid growth is desirable when it is paid for by loss of personal freedom, or even by bloodshed.

Furthermore, the economies of the Soviet Union and Eastern Europe turned in extremely poor growth performances in the 1970s and 1980s, while some of the fastest growth rates were recorded in the market economies of Japan, Taiwan, and Hong Kong. All told, history suggests that no other system can match market incentives for sustaining growth over long periods of time. Indeed, China has become the fastest-growing major economy on earth as it has moved away from planning toward markets (see pages 908–910).

Business fluctuations. A market economy is subject to periods of boom and bust, to inflation and unemployment. This holds true not only in capitalist market economies like the United States, but also in socialist market economies. So even though Marx dubbed the business cycle one of the fundamental flaws of *capitalism*, it is really a problem for *market economies*, be they capitalist or socialist.

Business fluctuations are not much of a problem in highly planned economies because total spending is controlled by the planners and is not permitted to get far out of line with the economy's capacity to produce. So the business cycle was not traditionally a serious problem in, say, the former Soviet Union or China. However, now that these two nations are transforming themselves into market economies, macroeconomic issues have become critical.

THE MARKET OR THE PLAN? THE SCOREBOARD

As we look back over this list, what do we find?

- Concerning *what to produce*, adherents to Western values will give a clear edge to the market, though conceding the need to curb some of its more flagrant

abuses. And, apparently, this view has now swept the former Soviet bloc as well.

- As to *productive efficiency*, almost everyone now seems to recognize the superiority of the market mechanism—as the opening quotations of this chapter suggest.
- But when we consider the *distribution of income*, we find that all societies have decided to plan; they differ only in degree.
- *High growth*, it seems, can be achieved—at least temporarily—with or without planning. Here an advanced nation will pause to question whether faster is always better, especially if growth comes from hard work and forced saving rather than through technological innovation. But in the poorer nations, the goal of rapid development is typically of paramount importance. Years ago, the urgent need for growth often led these countries toward central planning. Nowadays, however, more and more of them are becoming convinced that market systems, not planning, offer the best hope for long-run growth.
- Finally, in managing *business fluctuations*, there is no question that planned economies can do much better.

The results of our scoreboard are not entirely one-sided. But they do on balance point clearly toward market solutions—which is one reason why one country after another has switched from planning to markets. In sum:

The market has both strengths and weaknesses. Different countries—with their different political systems, value judgments, traditions, and aspirations—will draw the boundaries between the plan and the market in different places. But it now appears that almost all countries believe that the market mechanism should bear the primary burden of deciding which products to produce and how best to produce them.

CAPITALISM OR SOCIALISM?

Although the choice between capitalism and socialism seems to excite more ideological fervor, many observers believe it is less important than the choice between the market and the plan.

If it could design an appropriate incentive structure, a socialist *market* economy could do just as well as a capitalist market economy in terms of producing the right set of goods in the most efficient way. However, we have emphasized the word *if* to underscore the fact that designing such an incentive system may be quite difficult under socialism, even when markets are free.

Lacking the profit motive, a socialist society must provide incentives for its plant managers to perform well. This has proved difficult enough. But a still deeper problem is the need to maintain inventiveness, innovation, and risk-taking in an ever-changing world. Socialist nations, in which large accumulations of personal wealth are impossible, are noticeably low on "high rollers." This, many people feel, is why they were left behind in the 1980s.

Income distribution under socialism is naturally more equal than under capitalism simply because the profits of industry do not go to a small group of stockholders but instead are dispersed among the workers or among the populace as a whole. However, if supply and demand rules the labor market, a socialist nation may have as much inequality in the distribution of labor income as a

capitalist economy does—and for the same reasons: to attract workers into risky, or highly skilled, or difficult occupations.

The capitalist-socialist divergence is much more pronounced in regard to the issue of economic growth. To oversimplify, under capitalism it is the capitalists who determine the growth rate, while under socialism it is the state. Still, government incentives can prod capitalists to invest more; and socialist bureaucrats may squander resources on ill-advised investments. Examples of both fast and slow growth can be found under both systems.

Finally, as we have said, the severity of business fluctuations in a country depends much more on whether its economy is planned or unplanned than on whether its industries are publicly or privately owned.

SOCIALISM, PLANNING, AND FREEDOM

There is, however, a *noneconomic* aspect that is of the utmost importance in choosing between capitalism and socialism, or between the market and the plan, an issue whose importance was dramatized by the collapse of communism in Europe: *individual freedom.*

Planning must by necessity involve some degree of coercion; if it does not, then the plan may amount to little more than wishful thinking. In the extreme case of a command economy (such as Stalinist Russia or Nazi Germany), the abridgment of personal freedom is both painful and obvious. Less rigid forms of planning involve commensurately smaller infringements of individual rights, infringements that many people find tolerable.

Even in a market economy, some activities may be banned—such as prostitution and selling liquor to minors. Other economic activities may be compelled by law; safety devices in automobiles and labeling requirements on foods and drugs are two good examples. Each of these is planning of sorts, and each limits the freedom of some people. Yet most of these restrictions command broad public support in the United States.

Taxation is a still more subtle form of coercion. Many people do not view taxes as seriously impairing personal freedom because citizens remain free to choose the courses of action that suit them best. Indeed, this is one major reason why economists generally favor taxes over quotas and outright prohibitions. However, taxation can be a potent tool for changing individual behavior. As Chief Justice John Marshall pointed out with characteristic insight, "The power to tax involves the power to destroy."

Individual freedom is also involved in the choice between capitalism and socialism. After all, socialism imposes many more restrictions on what a person can do with his or her wealth. On the other hand, the poorest people in a capitalist society may find little joy in their "freedom" to go homeless and hungry.

Once again, it would be a mistake to paint the issue in black and white. Under rigid, authoritarian planning, the restrictions on individual liberties are so severe that they are probably intolerable to most people with Western values. That, presumably, is why many people in Eastern Europe reacted so strongly to the whiff of freedom. But more moderate and relaxed forms of planning—such as in some of the "Asian tigers"—seem compatible with personal freedoms.

Similarly, a doctrinaire brand of socialism that bans all private property (even the clothes on your back?) would entail a major loss of liberty. But a country with a large socialized sector can be basically free; the French, for example, do not feel notably less free than do Americans. And citizens of some countries with

largely capitalist economies, such as South Korea or Taiwan, did not enjoy political freedom until quite recently.

The real question is not *whether* we want to allow elements of socialism or planning to abridge our personal freedoms, but by *how much*.

Just as your freedom to extend your arm is limited by the proximity of your neighbor's chin, the freedom to build a factory need not extend to building it in the midst of a residential neighborhood. Just as freedom of speech does not justify yelling "Fire!" in a crowded movie theater when there is no fire, freedom of enterprise does not imply the right to monopolize trade. So, not surprisingly, different societies have struck the balance between the market and the plan, and between socialism and capitalism, in different places.

Most nations of Eastern Europe and the former Soviet Union are still in the midst of a difficult transition from plan to market, and from socialism to capitalism. But we cannot hope to understand this complex process unless we first learn how the old, centrally planned system of the Soviet Union functioned.

THE ADMINISTRATIVE COMMAND ECONOMY: THE OLD SOVIET UNION

In November of 1917, a determined group of Bolsheviks led by V. I. Lenin overthrew a short-lived democratic government to establish a new government dominated by the Communist party. The Soviet Union was a relatively backward country then, and its economy was exhausted by a civil war. Seeing the new nation on the verge of both political and economic collapse by 1921, Lenin ushered in his New Economic Policy: a brief era of open discussion and dialogue in which markets returned and the economy prospered.

But his successor, Joseph Stalin, chose a very different—and costly—course when he rose to power. Starting in 1928, Stalin implemented one of the most brutal regimes of modern times. He made two major decisions. First, he forcibly introduced socialized agriculture in a process known as *collectivization*. Second, he nationalized the means of production and eliminated markets. *Central planning* was installed as the preeminent mechanism for allocating resources.

These events left a lasting mark on the Soviet economy. Indeed, several notable characteristics defined the Soviet system from Stalin's day right up to its demise in 1991:

- The economic system was *hierarchical* and *centralized*. The basic goals of economic activity were determined by the Communist party and left to be fulfilled by local enterprise managers.
- Stalinist economic policies stressed maximal economic growth, with emphasis on *heavy industry* and, especially, on the *military*. Consumer goods and agriculture were considered less important.
- The system was largely *quantity-directed*, with only limited use of the price mechanism.

CENTRAL PLANNING IN THE COMMAND ECONOMY

Let us see, briefly, how the old central-planning system worked. Ironically, the structure of the Soviet economic system resembled the hierarchy of a giant corporation. At the top was a single strongman or a ruling clique. The political leadership functioned like the chairman of the board, setting overall policy

objectives, but had much more absolute authority than the chairman of any corporation. Next came the State Planning Commission (*Gosplan*) which translated the goals and priorities established by the top echelons of the Communist party into Five-Year and One-Year Plans.

The *Five-Year Plans* set the nation's basic strategy for resource allocation: How much for investment? How much for military procurement and for scientific research? Would there be more cars or more refrigerators? Observers in both the West and the Soviet Union paid close attention to the numerical goals in these plans—and to the Soviet Union's failure to meet them.

The Five-Year Plans were too vague to serve as blueprints for action, however. This job was left to the *One-Year Plans*—enormous sets of documents that covered almost every facet of Soviet economic life. Planners translated broad national goals into specific directives for subordinate ministries, agencies, and regional authorities. But the One-Year Plans were so detailed and complex that they were rarely completed until well into the year, if then.

Several more layers of bureaucracy intervened before reaching the level of the enterprise managers, who acted more like bureaucrats than entrepreneurs in the Soviet system. Communications within the hierarchy were predominantly vertical. Orders flowed down from top to bottom, while data flowed up from bottom to top. Since the data requirements were immense and the number of layers within the bureaucracy so large, the problems of accurate data transmission and processing were monumental. It is precisely this problem, of course, that the market mechanism solves so neatly: Prices convey most of the information that anyone needs.

THE ROAD TO *PERESTROIKA*: WHY DID THE SOVIET ECONOMY FAIL?

The economic system that we have just briefly described was plagued by many problems. Together they stunted productivity growth and consequently slowed the growth of output in the Soviet Union. While it is difficult to pinpoint the precise causes of the Soviet growth slowdown, several contributing factors are clear.

First, while central planning was powerful, it was also crude. It did not adapt well to either changing technology or the increasing complexity of modern economic life. The system, in a word, became ossified.

Second, Soviet planning emphasized high levels of output, with little attention to issues like product quality, low cost, and catering to consumer desires. Because managerial bonuses depended on fulfilling the plan, managers would do dysfunctional things to meet their quotas. For example, scarce materials were routinely hoarded, thus worsening chronic shortages. Or, if shoe leather was in short supply, factory managers might produce only children's shoes rather than adult shoes.

Third, the Soviet system inhibited change, since change threatened fulfillment of the plan. Even when new technology could be imported, diffusion of these new ideas often did not take place. There were simply no incentives to innovate.

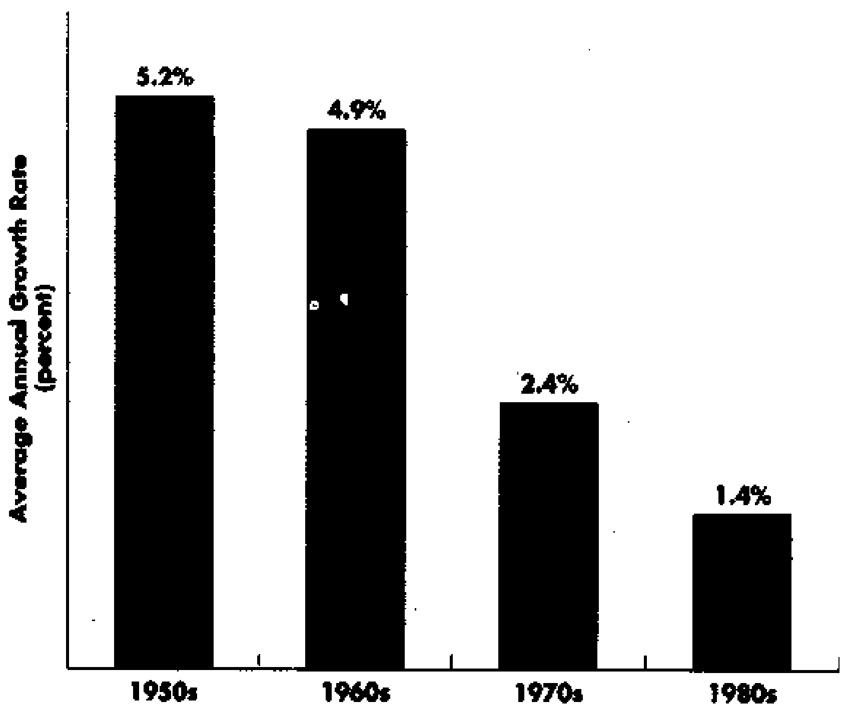
Fourth, as time passed, workers became increasingly irritated that their earnings were not translated into better consumer goods and improved standards of living. (Refer back to the first opening quotation of this chapter.) Modern telecommunications exacerbated this problem by enabling Soviet workers to learn about the goods and services enjoyed in the West.

Attempts to reform the command economy date back to the Khrushchev era of the 1950s. But they were almost always superficial in character—designed to sustain the basic features of the planning system. And they were always imple-

FIGURE 38-1 THE SLOWDOWN OF SOVIET ECONOMIC GROWTH

This chart shows that rapid Soviet economic growth in the 1950s and 1960s dwindled rapidly in the 1970s and 1980s.

SOURCE: Central Intelligence Agency.



mented on a limited scale. As a result, several waves of "reform" had little effect on the performance of the Soviet economy. Thus, from the 1960s onward, the rate of growth of productivity and output declined almost continuously. (See Figure 38-1.)

In 1985, Mikhail Gorbachev generated genuine excitement when he assumed power and announced both *perestroika* (restructuring) and *glasnost* (openness). Each seemed to call for fundamental changes in the Soviet system. Cast against the authoritarian background of the Soviet Union, the concept of *glasnost*, which involved the expansion of democratic ideas in both the political and economic spheres, seemed to be a major departure. Moreover, combined with *perestroika*, which involved major changes in the Soviet economy, ranging from less planning to greater managerial autonomy to decentralization of foreign trade decisions, the Gorbachev concept of radical reform seemed to be an appropriate prescription.

But in the end, *perestroika* failed because, in a sense, it was another partial reform designed to shore up the shaky socialist edifice. Moreover, as with previous attempts at reform, implementation encountered fierce resistance from the entrenched bureaucracy. Only when Boris Yeltsin asserted his authority and made clear his support for the transition to markets did the old Soviet system come to an end.

ECONOMIES IN TRANSITION: RUSSIA AND HER NEIGHBORS

The demise of central planning left the former Soviet Union and much of Eastern Europe with a daunting task: to build a system of capitalism and markets atop the wreckage of socialist central planning. In the West, market capitalism

as we know it evolved gradually over hundreds of years; no one ever designed it. So there was no instruction book that the leaders of the formerly socialist countries could consult for advice about the transition process.

Not surprisingly, each country followed a different route, and experiences since the 1989 to 1991 period have differed greatly. Some countries, like the Czech Republic and Poland, decided to move as quickly as possible. The Czechs, in particular, handled the transition remarkably smoothly and are now mostly through it. Others, like Hungary and especially Russia, took a more cautious and piecemeal—some might say, uncertain—approach. Russia's prodigious economic problems were (and to some extent still are) complicated by considerable political instability and a breakdown of law and order. So progress was slow at first, and has been quite stressful; their transition is far from over. Nonetheless, despite these differences, the transition processes in the formerly socialist countries have had some common elements:

PRIVATIZATION means transferring publicly owned property to private hands.

1. *Privatization.* If socialism (public ownership) is to give way to capitalism (private ownership), the government must find some way to place assets in private hands—a process known as **privatization**. This was no easy task in a country like Poland or the Czech Republic, where the state had confiscated the means of production from individual owners decades ago. And it was even harder in Russia, which never really experienced capitalism.

   To establish a viable system of private ownership, several major questions needed to be addressed. How were ownership "rights" to be identified and parceled out? How could assets be valued in a country without a functioning price system? What was to be done about loss-making enterprises—businesses that might well have negative value under capitalism?[1] Where would the necessary business managers be found?

   In Russia, privatization proceeded slowly and unevenly at first, but has lately picked up speed. Some assets were directly distributed—for example, by giving shares to workers and apartments to tenants. Others were distributed by variants of the famous voucher system pioneered in what was then Czechoslovakia. Under this system, citizens are given a kind of fictitious money ("vouchers"), which they use to bid for shares in companies being sold at auction. So many companies were privatized in this way that Russia now has more citizen-shareholders than the United States!

2. *Establishing markets and prices.* Developing a price system virtually from scratch is also complicated. Most planned, socialist economies had made only minimal use of prices for allocating resources. So the prices that existed when the planning systems ended bore little resemblance to market prices and, in particular, did not reflect relative scarcities. For example, bread was priced so cheaply in the old Soviet Union that farmers used it instead of wheat to feed their cows.

   Although it was agreed that privatization would eventually lead to the development of markets, the process would take time. In the meantime, something had to be done about the inappropriate prices that existed. Those that had previously been held at ludicrously low levels were freed from control, though in varying degrees in different countries. The predictable result in most countries was inflation, which sharply reduced the real standards of living of most people.

---

[1] In one famous case, a major Italian auto company was offered an East German automobile factory for free. It turned the "gift" down!

3. *Macroeconomic stabilization.* As noted earlier in this chapter, neither inflation nor unemployment were serious problems in the planned economies. In consequence, few of them had ever used monetary or fiscal policy in the manner typical of market economies. This led to some serious problems when the old systems began to disintegrate. As just mentioned, inflation leaped upward in most transition economies. Government budget deficits increased because revenues could not be maintained with inappropriate tax structures and collapsing political systems.

   The large state enterprises posed a particularly cruel dilemma. If they were cut adrift to fend for themselves in world markets, most would have failed, leading to mass unemployment. None of the formerly socialist nations were willing to do this. Instead, subsidies to many loss-making enterprises were maintained—which widened government budget deficits. Since there were no financial markets as we know them, large budgets led to rapid expansion of the money supplies—which, of course, fueled inflation.
4. *Disruption of international trade.* One of the major features of the planned socialist systems was central control of international trade. In practice, they traded mostly with one another. When the planning system collapsed, trade decisions had to be decentralized. Worse yet, with the economies of its trading partners contracting rapidly, each country found its traditional export markets shrinking.

Any one of these four problems, by itself, would pose a major hurdle for an advanced Western economy. For example, the United Kingdom under Margaret

A TALE OF TWO COUNTRIES

After the Berlin Wall fell, the formerly socialist nations scrambled to implement economic reforms. The ensuing 6 years witnessed both stunning success stories and colossal debacles as nations long accustomed to command economies struggled to cope with free markets.

POLAND: TWO STEPS FORWARD, ONE BACK

No former Soviet satellite embraced economic reform more firmly in the late 1980s than Poland. But by 1995, President Lech Walesa, once a national icon, had lost a reelection bid to a former communist apparatchik. Walesa's repudiation was not, however, a signal that voters were ready to return to a planned economic system. Ironically, public opinion polls suggested that Poles were growing increasingly impatient with the slow pace of reform!

Nonetheless, reformers could find ample cause for concern. New laws stretched out the timetable for privatizing state-owned enterprises. And industries deemed "strategic," such as energy and telecommunications, could not be privatized without parliamentary approval. The state also began flexing more regulatory muscle. Free-market purists were worried that, when privatization resumes, political cronyism may count for more than economic efficiency.

THE CZECH REPUBLIC: FULL SPEED AHEAD

The Czech Republic adopted capitalism seemingly without breaking stride. Mass privatization there is now nearly complete. In 1995, the economy was growing at a steady 4 percent clip, with inflation at 10 percent and falling, minuscule unemployment, and a budget surplus.

The Czech success story combined both luck and skill. The Civic Democratic Party (CDP) of Prime Minister Vaclav Klaus, an ardent free marketeer, has assumed popular control. And there has yet to be any serious suggestion of returning to socialism, despite the fact that 40 percent of the people consider themselves poor.

Top CDP officials have taken a hands-on approach to ensuring that the free-market playing field remains level by rooting out corrupt officeholders and identifying conflicts of interests.

SOURCE: "Into the Trough," *The Economist*, July 15, 1995, p. 38; Casey Bukro, "Eyes Are Looking East to Former Soviet Bloc," *Chicago Tribune*, July 12, 1995, p. 1; Bob Williams, "Looking East for Salvation," *The News & Observer*, April 10, 1995, p. A1; and "Paradox Explained," *The Economist*, July 22, 1995, p. 52.

Thatcher took years to privatize a relatively small share of its economy—and made some mistakes along the way. Now try to imagine tackling all four problems—plus others we have not even mentioned—at once, and doing so in a nation that has just gone through a political revolution. That was the task facing the economies in transition. Every one of these countries suffered a severe contraction of output that lasted several years. By now all are growing again; but the transition process is far from over.

CHINA UNDER MAO: REVOLUTIONARY COMMUNISM

The People's Republic of China makes a good case study for this chapter because it has spent nearly 50 years experimenting with different economic models. In the process, the Chinese have vividly confronted the fundamental questions of this chapter. Socialism or capitalism? Market or plan? And they have come up with different answers at different times.

After the communist takeover in 1949, the Chinese economy was patterned on the Russian model. It was a Soviet-style command economy with emphasis on rapid economic growth, particularly industrial growth. But there were important differences. Probably the most important of these was the decision by Mao Tse-tung *not* to rely on *material incentives* to motivate the work force. Mao and the Chinese leadership were disdainful of such "bourgeois" motives; they preferred exhortation, patriotism, and, where necessary, force.

Russian communism bent socialist doctrine to accommodate human nature. But Chinese communism at first insisted on bending human nature to accommodate Maoist doctrine—to create "the new man in the new China," an effort that was abandoned more than 20 years ago.

A second, and more lasting, difference is that China was always less centralized than Russia. This has been true for centuries and, under central planning, provincial authorities in China had more power and discretion than they did in the Soviet Union. Even today, the leaders in Beijing often have trouble controlling the provinces.

Chinese economic growth has proceeded in fits and starts.

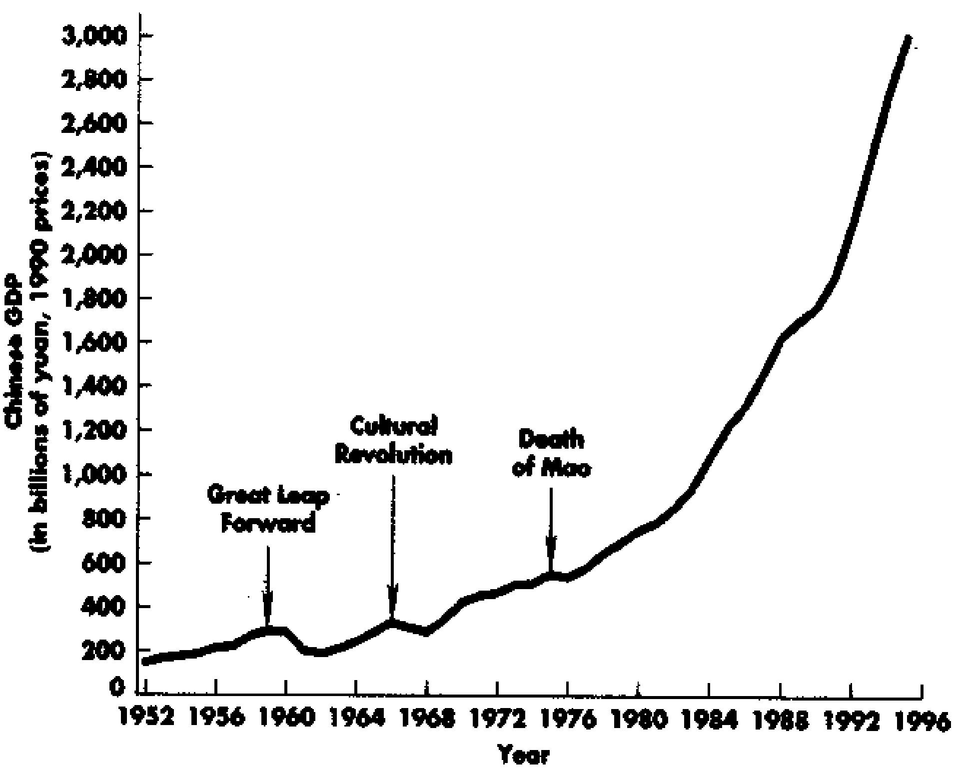
The immediate problem after the Maoist takeover was to lift China out of wartime devastation and to establish communist institutions and values in a vast and semiliterate land. With Soviet assistance, the plan succeeded. But then Mao changed course.

China's next step, the *Great Leap Forward* (1958–1960), turned out to be a giant step backward. The Maoist regime set unrealistically ambitious production goals and selected means for implementing them that were more romantic than rational. Material incentives were deemphasized. Ideologically pure "Reds" were supposed to replace technocratic "experts." Massive applications of brute labor were supposed to compensate for shortages of machinery and advanced technology.

It was a pipe dream or, more accurately, a nightmare for China. National income fell substantially and took years to recover (see Figure 38-2). The ideological excesses of the period also raised tensions with the Soviet Union.

Finally, the Chinese leadership threw out the "Reds" and brought back the "experts," signaling a return to rational economic calculation. The early 1960s were marked by a return to the Soviet planning model, though more decentralized. Economic growth resumed.

FIGURE 38-2 REAL GDP IN THE PEOPLE'S REPUBLIC OF CHINA[a]



While Chinese economic statistics are notoriously inaccurate, the official data are portrayed here. The Great Leap Forward and the Cultural Revolution stand out as major blemishes on the record of Chinese economic growth. Growth has been extremely rapid since the 1980s.

[a]Data are the net material product for 1952–1977 and real GDP for 1978–1995.

SOURCE: *International Financial Statistics* and *China Statistical Yearbook*.

Then, inexplicably, Mao changed China's course once again with the *Great Proletarian Cultural Revolution* (1965–1969). The "Reds" were in and the "experts" were out as never before. The infamous Red Guards were sent out to purge rightist elements from Chinese society, organize revolutionary cadres, and spread the teachings of Chairman Mao. If anyone worried about economic productivity in this environment, it did not show. By the summer of 1967, both the Chinese economy and other elements of Chinese society were in disarray. Output fell again, but recovered more quickly than it had from the Great Leap (see again Figure 38-2).

Finally, in Mao's last years, Chinese revolutionary fever receded and the "experts" were rehabilitated once again. Material incentives and rational economic calculation returned to China, as did economic growth.

CHINA SINCE MAO: EXPLOSIVE GROWTH

The leaders who succeeded Mao were less doctrinaire about their communism and far more interested in results. China began, cautiously at first, to step away from the Soviet model and adopt important features of the market economy in its place. It welcomed Western tourism, trade, and technology. Chinese managers, engineers, and economists went to study in the West. Material incentives

were emphasized and even small-scale capitalism was allowed back on the Chinese mainland.

These trends accelerated dramatically in the 1980s and 1990s, as market forces were increasingly allowed to supplant central planning. Farmers were given land to do with as they pleased. Private shops and other small businesses flourished. Foreign companies were invited to set up operations in China, and many did. China became the biggest Asian growth miracle of them all. Between 1981 and 1992, economic growth averaged almost 10 percent per year—a stunning performance that raised standards of living tremendously.

By the mid-1990s, China was still burdened with many inefficient, state-owned enterprises. And the Communist party, both scarred and scared by the Tiananmen Square revolt in 1989, still held a firm grip on the political system. But Western visitors to China found it difficult to see much evidence of socialism. Cities like Beijing, Shanghai, and Guangzhou gave every appearance of being hustling, bustling market economies on the make.

JAPAN'S UNIQUE BRAND OF CAPITALISM

But China has a long way to go before it even comes close to matching the economic achievements of Japan, which is certainly *the* outstanding success story of the period since World War II. From 1955—when Japan began its productivity enhancement campaign—to 1989, real GNP in Japan rose by an astounding 836 percent.[2] (The corresponding expansion of the U.S. economy was only 177 percent.) Its automobile, electronics, and semiconductor industries, to name just a few, lead the world in technological and manufacturing prowess. Its banks and other financial institutions are the world's largest.

How did Japan do it? It is true, but far too simple, to say that Japan succeeded by adopting free-market capitalism—for both free markets and capitalism look quite different in Japan than in the United States. We devote most of the rest of this chapter to describing some of these differences.

EXPORT-LED GROWTH

Japan is a crowded, island nation far removed from the world's major markets. Almost totally devoid of natural resources, it must import large amounts of raw materials, energy, and foodstuffs just to survive. And yet, by concentrating on manufacturing and exporting, it has managed not just to prosper but to propel itself into the forefront of nations.

EXPORT-LED GROWTH refers to the strategy of emphasizing the production of goods for export.

One secret to Japanese economic success has certainly been its emphasis on both high levels of investment and **export-led growth**. From 1955 to 1989, Japanese exports in real terms grew an astounding 3,227 percent. Indeed, Japan's export machine has been so successful that its huge trade surplus has been the cause of international trade frictions for years. (See the box on the next page.)

There is no single secret to Japan's exporting success. High levels of investment have certainly facilitated the adoption of the latest technology. Japanese industry is clearly outward looking in a way that American industry is not. Many organizations, such as the fabled Ministry of International Trade and Industry (MITI), work to promote exports. Japanese business has also shown a remarkable ability to adapt to changing world markets. As one industry (say, ship-

[2] We stop the comparison in 1989 because Japan has experienced a protracted recession in the 1990s.

U.S.-JAPANESE TRADE: IS EVERYONE PLAYING FAIR?

Japan's huge trade surpluses with the United States have threatened friendly relations between the two countries. Several rounds of trade negotiations during the Bush and Clinton administrations have reduced the surplus only slightly. Why? One former trade negotiator believes that leaders of both nations incorrectly assume that the Japanese and U.S. economies are organized in the same way.

> As long as this assumption is accepted, there are only two possible explanations for our problems: Someone is cheating or someone is performing poorly....
>
> There is a third explanation: The two sides do not share the same economic views. The U.S. runs its economy for the consumer, prevents or strictly regulates aggregation of corporate power, eschews industrial policy, views the government role as that of referee and calls for a laissez-faire brand of free trade. Japan emphasizes the interest of the producer, promotes agglomeration of corporate power, pursues an active industrial policy, sees the government as a coach and practices a mercantilistic brand of free trade.
>
> In effect, Japan is playing football while the U.S. is playing baseball. Japan is not playing unfairly, and the U.S. team is trying as hard as it can. But football is a rougher game than baseball, and the baseball players are taking a beating. They will continue to do so as long as their leaders insist that the games are really the same.

SOURCE: Clyde Prestowitz, "George and Toshiki—The Odd Couple," *The New York Times*, March 10, 1990.

building) declines, the Japanese shift into another (say, consumer electronics) whose star is rising. Above all, the major Japanese companies seek always to grow—and that means exporting.

"JAPAN, INC."

Unlike the United States, but like much of Western Europe, Japan has no tradition of enforcing antitrust laws. Indeed, at times the government has seemed to promote rather than oppose bigness, perhaps as a way to catch up to the West. In consequence, Japanese industry is far more concentrated than American industry, especially in manufacturing and financial services. Industrial concentration, barriers to imports, inefficient retailing, and an expensive currency combine to keep domestic prices of consumer goods astoundingly high.

The cozy relationship between the big Japanese corporations and the Japanese government has acquired the nickname "Japan, Inc.," to indicate that economic policy there is geared more to producer than to consumer interests. Indeed, for decades Japan was said to be ruled by an "iron triangle" consisting of giant corporations, powerful career bureaucrats, and politicians from the dominant Liberal Democratic Party (LDP).

Lately, however, this system has shown unmistakable signs of strain. The severe and protracted recession of the 1990s weakened many Japanese companies, especially the banks. The government's inability to set things right tarnished the reputations of the prestigious ministries. And then, in 1993, the LDP lost control of the government for the first time since its ascendancy in the 1950s. As this text is written, both political reform and financial reform stand high on the Japanese agenda.

JAPANESE MANAGEMENT METHODS

The internal organization of a Japanese company also looks quite different from an American corporation.[3] An industrial giant like Toyota may be surrounded

[3] However, in recent years, more and more American companies have adopted Japanese management techniques.

by satellite companies that supply parts and help it operate its famous "just-in-time" inventory system, a highly efficient way of organizing the factory floor to minimize delays. These smaller companies live at the mercy of the corporate giants and act as shock absorbers when demand declines.

In addition, members of Japan's manufacturing combines (*keiretsu*) own chunks of each other's stock and forge tight links with Japanese banks, the world's largest. This gives Japanese industry access to cheap, "patient" capital that is willing to wait for a return on its investment. Japanese managers pay scant attention to daily movements of the stock market and never worry about hostile takeovers. All this helps Japanese industry maintain its long-run focus.

Japan also differs from the United States in the way its work force is organized and paid. For one thing, many employees of large Japanese corporations have *lifetime employment* guarantees and are never laid off. These features help align the interests of labor and management, build loyalty to the company, and make Japanese workers less resistant to change than their American counterparts. For example, if a Japanese plant automates, most employees know they will not only keep their jobs but share in any gains automation may bring.

The Japanese workplace is also less hierarchical. Pay differentials between executives and production workers are a fraction of those in the United States; so, naturally, the Japanese distribution of income is much more equal. Managers eat in the same cafeterias, drink in the same bars, and sometimes even wear the same uniforms as blue-collar workers. Japanese workers are also consulted closely on how the factory is to be run. Decision making is by consensus—even if consensus takes a long time to build. American management, by contrast, exercises more top-down control.

For these reasons, and perhaps also because of strong conformist tendencies within Japanese society, labor-management relations in Japan are less adversarial and more cooperative than those in the United States or Europe.

IS JAPAN A PLANNED ECONOMY?

All this did not happen by the invisible hand alone. The system was consciously designed to raise industrial productivity, keep it growing fast, and turn Japan into an industrial powerhouse that would rival the Western nations. Many Japanese innovations, such as "just-in-time" manufacturing, originated in the private sector. But elite government bureaucrats played an active role in promoting what they saw as good ideas, discouraging bad ones and, in general, lending a helping and guiding hand.

The role of the government, and especially MITI, in Japan's economic success is highly controversial. Ardent free-marketeers downplay its contribution and point to episodes in which MITI clearly got in the way—such as its ill-conceived attempt to drive Honda out of the automobile business in the 1960s. Protectionists seeking to limit Japanese imports exaggerate the role of MITI and portray Japan, Inc., as a monolith, which it certainly is not. In their eyes, Japan's *industrial policy* is the key to its success.

A balanced view of the matter seems hard to strike, especially since few Westerners really fathom the clubby ways of Japanese businesses. But a few things seem clear. First, the Ministry of Finance and the Bank of Japan (the central bank) have exercised far more comprehensive control over the Japanese financial system than the Treasury and the Federal Reserve have ever done here. Second, Japanese industrial policy, while far from infallible, seems to have had consid-

CAN AMERICA IMPORT JAPANESE MANUFACTURING METHODS?

Most [Japanese] firms [in the United States] have adopted something like joint consultation and related Japanese practices, including an elaborate screening of job applicants. . . . [S]ome U.S. workers who were hired at Honda in Ohio were reportedly surprised both at the many questions they were asked that seemed unrelated to work and at the length of the interviews, which were attended by executives and vice presidents. . . . Once hired, these workers reported attending frequent meetings with the management on production matters. Such frequent meetings at the Honda plant in Ohio are signified by the slogan, "let's Y-gaya," which means in fractured Japanese, "let's have a bull session." In these operations management and workers share the same tables for lunch, thereby creating an informal setting for communication. . . .

Productivity at Nummi [a joint Toyota-General Motors factory in Fremont, California] after only one year of operation was reported to have increased by 48.5 percent over what it was at the old Fremont plant under GM management. Absenteeism and drug use, which plagued the old Fremont plant, dropped dramatically after Nummi took over. Nummi's efforts at productivity enhancement continue with the slogan "let's kaizen," or "let's improve." Also, in contrast to the old Fremont plant, the quality of the automobiles produced at Nummi has been rated highly. . . .

. . . Honda in Japan is known for its emphasis on nurturing the sense of cooperative teamwork among its workers. This emphasis was imported to Honda's operation in Ohio. At the Ohio plant workers, referred to as *associates*, are encouraged to acquire skills and training by continual interactions with one another on the shop floor rather than through formal training sessions. Productivity at the plant reportedly approaches that of Honda's plants in Japan, and the quality of the automobiles produced in Ohio is said to equal that of Japanese-made Hondas.

SOURCE: Masanori Hashimoto, "Employment and Wage Systems in Japan and Their Implications for Productivity," in Alan S. Blinder (ed.), *Paying for Productivity: A Look at the Evidence* (Washington, D.C.: Brookings Institution, 1990), pp. 292–293.



erable success in assisting winners and shutting down losers. Third, Japanese industry benefits enormously from a work force that may be the best educated and most cooperative in the world; this is certainly a substantial achievement of government.

WHAT CAN THE UNITED STATES LEARN FROM (AND TEACH) JAPAN?

Learning is a two-way street. The Japanese have learned much from American industry and continue to do so; their famed "quality circles," for example, orginated here. Economywide productivity continues to be higher in the United States than in Japan. Nonetheless, the Japanese seem to be the technological leader in some important industries—such as automobiles and robotics. There,

we can "play catch-up" by observing them, just as they did years ago by observing us.

Many people feel, however, that the most important things we can learn from the Japanese are not the latest innovations in robotics or chip manufacturing, but rather their ways of organizing and motivating *people*. The process begins with an education system that has virtually abolished illiteracy, continues through Japan's unique labor-relations system, and includes production and management techniques that are being emulated all over the world. (See the box on the previous page for some examples of importing Japanese management methods to America.)

Some observers predict that, as the Japanese get richer, they will come to behave more like Americans. Already, for example, there are signs that Japanese households are saving less, Japanese workers are changing jobs more frequently, and there is strong sentiment to "deregulate" their economy somewhat along American lines. Others, however, stress basic cultural differences between the two societies that change very slowly, if at all. The two brands of capitalism are indeed different. Only time will tell which system adapts more successfully to our ever-changing world.

SUMMARY

1. Economic systems differ in the amount of planning they do and in the extent to which they permit private ownership of property. However, **socialism** (state ownership of the means of production) need not go hand in hand with central planning, and **capitalism** need not rely on free markets. The two choices are distinct, at least conceptually.
2. *Free markets* seem to do a good job of selecting the bill of goods and services to be produced and at choosing the most efficient techniques for producing these goods and services. Planned systems have difficulties with both of these choices.
3. Market economies, however, do not guarantee an equitable distribution of income and are often plagued by business fluctuations. In these two areas, *planning* seems to have some advantages.
4. A major problem for socialism is how to motivate management to achieve maximal efficiency and to maintain inventiveness in the absence of the profit motive.
5. Individual freedom is a noneconomic goal that is of major importance in the choice among economic systems. Any element of planning or of socialism infringes upon the freedoms of some individuals. Yet complete freedom does not exist anywhere, and certain limitations on individual freedom command wide popular support.
6. From the days of Stalin until its dissolution in 1991, the Soviet Union followed a rigid system of *central planning* which emphasized heavy industry and armaments and deemphasized consumer needs.
7. Planning in the Soviet Union was bureaucratic and hierarchical and encountered monumental difficulties in adapting to modern technology, motivating workers and managers, and achieving satisfactory productivity.
8. While Soviet consumers were free to spend their money as they pleased, there was little **consumer sovereignty**. Planners, not consumers, decided what would be produced. The labor market, however, operated much as it does in America—using wage rates to equate supply and demand.
9. With the end of centrally planned socialism in Eastern Europe and the former Soviet Union, planned systems are being replaced by market systems, a difficult process generally described as *transition*.
10. This transition process involves difficult problems like privatizing assets, freeing prices from controls, stabilizing the macroeconomy, and reestablishing international trade.
11. The Chinese economic system has changed several times since the communist takeover in 1949, passing through several periods of intense revolutionary fervor and little economic progress. Planning there was similar to that in Russia, although somewhat less centralized.
12. Beginning in the late 1970s, the Chinese introduced important aspects of the market economy—and even elements of capitalism—into their economic system. Lately, Chinese growth rates have been the highest in the world.

13. Japan has used **export-led growth** to propel itself to the forefront of nations; but lately its single-minded concentration on exporting has been a source of international trade tensions.
14. Japan has more industrial concentration than the United States, a less adversarial system of labor-management relations, tighter links between manufacturing companies and banks, and more active cooperation between government and industry. Observers disagree about the relative importance of each of these influences in accounting for Japan's industrial success.

KEY TERMS

Capitalism
Socialism
Planning
Free markets
Consumer sovereignty
Collectivization
Perestroika
Privatization
Material incentives
Great Leap Forward
Great Proletarian Cultural Revolution
Export-led growth
Industrial policy

QUESTIONS FOR REVIEW

1. Explain why the choice between capitalism and socialism is not the same as the choice between markets and central planning. Cite an example of a socialist, market economy and of a planned, capitalist economy.
2. If you were the leader of a small, developing country, what are some of the factors that would weigh heavily in your choice of an economic system?
3. Which type of economic system generally has the most trouble achieving each of the following goals? In each case, explain why.
   a. An equal distribution of income
   b. Adequate incentives for industrial managers
   c. Eliminating business fluctuations
   d. Balancing supply and demand for inputs
4. If you were a plant manager under old-style Soviet planning, what are some of the things you might do to make your life easier and more successful? (Use your imagination. Russian plant managers did!)
5. What are some of the major issues involved in the transition from plan to market, and why is this transition so difficult in practice?
6. Do you think the U.S. government should take a more active role in guiding American industry? If so, is Japan a good model? If not, why is Japan not a good model?
7. During the early 1990s, Japan and the United States engaged in talks on "structural impediments" designed to achieve more balanced trade between the two nations. As part of these talks, U.S. government representatives told the Japanese how they could make their economy more like ours, while Japanese negotiators told their American counterparts how to make the United States more like Japan. Does this make sense to you? Who should be emulating whom?

13. Japan has used **export-led growth** to propel itself to the forefront of nations; but lately its single-minded concentration on exporting has been a source of international trade tensions.
14. Japan has more industrial concentration than the United States, a less adversarial system of labor-management relations, tighter links between manufacturing companies and banks, and more active cooperation between government and industry. Observers disagree about the relative importance of each of these influences in accounting for Japan's industrial success.

KEY TERMS

Capitalism
Socialism
Planning
Free markets
Consumer sovereignty
Collectivization
Perestroika
Privatization
Material incentives
Great Leap Forward
Great Proletarian Cultural Revolution
Export-led growth
Industrial policy

QUESTIONS FOR REVIEW

1. Explain why the choice between capitalism and socialism is not the same as the choice between markets and central planning. Cite an example of a socialist, market economy and of a planned, capitalist economy.
2. If you were the leader of a small, developing country, what are some of the factors that would weigh heavily in your choice of an economic system?
3. Which type of economic system generally has the most trouble achieving each of the following goals? In each case, explain why.
   a. An equal distribution of income
   b. Adequate incentives for industrial managers
   c. Eliminating business fluctuations
   d. Balancing supply and demand for inputs
4. If you were a plant manager under old-style Soviet planning, what are some of the things you might do to make your life easier and more successful? (Use your imagination. Russian plant managers did!)
5. What are some of the major issues involved in the transition from plan to market, and why is this transition so difficult in practice?
6. Do you think the U.S. government should take a more active role in guiding American industry? If so, is Japan a good model? If not, why is Japan not a good model?
7. During the early 1990s, Japan and the United States engaged in talks on "structural impediments" designed to achieve more balanced trade between the two nations. As part of these talks, U.S. government representatives told the Japanese how they could make their economy more like ours, while Japanese negotiators told their American counterparts how to make the United States more like Japan. Does this make sense to you? Who should be emulating whom?